Multivariate Dichteschätzung in der explorativen Datenanalyse

Dissertation

zur Erlangung des akademischen Grades

Dr. rer. nat.

vorgelegt der

Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Augsburg

von Hao Wang

03.11.2009

Wissenschaftliche Betreuung: Prof. Dr. Antony Unwin

1. Gutachter: Prof. Dr. Antony Unwin

2. Gutachter: Prof. Dr. Friedrich Pukelsheim

Tag der mündlichen Prüfung: 27.01.2010

Danksagung

Die vorliegende Dissertation wurde am Lehrstuhl für rechnerorientierte Statistik und Datenanalyse der Universität Augsburg angefertigt.

An dieser Stelle möchte ich mich ganz herzlich bei all denen bedanken, die mir während meines ganzen Studiums in Augsburg mit Rat und Tat zur Seite standen und mich auf unterschiedliche Weise unterstützt haben.

Allen voran gilt der Dank meinem Doktorvater, Herrn Prof. Dr. Antony Unwin für die Vergabe des interessanten Promotionsthemas, die fachliche und zielgerichtete Betreuung, die moralische Unterstützung sowie sein menschliches Engagement bei dieser Dissertation.

Für die Übernahme des Zweitgutachtens meiner Dissertation möchte ich mich auch recht herzlich bei Herrn Prof. Dr. Friedrich Pukelsheim bedanken.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einführung

> Die statistische Methode ist eines jener Verfahren, die der menschliche Geist zum Zwecke der Erkenntnis des Lebens und der Lebensverhältnisse der Menschen wie zur Erkenntnis der den Menschen umgebenden Natur ausgebildet hat.
>
> Franz Zizek (1938)

In der statistischen Datenanalyse wird üblicherweise angenommen, dass die empirischen Daten einer Wahrscheinlichkeitsverteilung unterliegen. Das Ziel der Dichteschätzung besteht darin, die unbekannte Dichtefunktion dieser unterliegenden Wahrscheinlichkeitsverteilung anhand der empirischen Beobachtungen mittels statistischen Verfahren zu schätzen. In der explorativen Datenanalyse kann die in den Daten versteckte nützliche Information aufgrund der geschätzten Dichte untersucht werden.

Grob teilt man Dichteschätzung in parametrische Dichteschätzung und nichtparametrische Dichteschätzung. Die parametrische Dichteschätzung geht auf Fisher (1922, 1932) zurück. Fisher schlug die folgenden zwei Phasen für die Dichteschätzung vor:

- „Specification. - Problems of specification are those in which it is required to specify the mathematical form of the distribution of the hypothetical population from which a sample is to be regarded as drawn.“ und
- „Estimation. - Problems of estimation are those in which it is required to estimate the value of one or more of the population parameters from a random sample of the population.“

Nichtparametrische Dichteschätzung geht auf Pearson (1893, 1902) (vgl. Scott (1992)) zurück, wobei die unbekannte Dichtefunktion ohne Verteilungsannahme direkt aus den empirischen Daten zu schätzen ist. Heutzutage spielt nichtparametrische Dichteschätzung in der explorativen Datenanalyse mit der Entwicklung der Computertechnik eine immer wichtigere Rolle, weil parametrische Dichteschätzung stark von der Modellannahme abhängt und deswegen in der Praxis nicht einwandfrei anwendbar ist, falls man wenig a priori Information über das unterliegende Modell hat. Die vorliegende Arbeit konzentriert sich auf eine in der Praxis oft eingesetzte Familie der nichtparametrischen Dichteschätzung, nämlich die Kerndichteschätzung, um die Anwendung der nichtparametrischen Dichteschätzung in der explorativen Datenanalyse zu erläutern. In dieser Arbeit wird die Gaussian Kernfunktion als Default bei Kerndichteschätzung verwendet.

Die Auswahl des Glättungsparameters hat einen entscheidenden Einfluss auf Kerndichteschätzung. In den letzten Jahren wurde eine Vielzahl von statistischen Verfahren entwickelt, um einen in Bezug auf ein gewisses Kriterium optimalen Glättungsparameter bei Kerndichteschätzung zu bestimmen. Eine interessante Zusammenfassung dieser Methoden findet man in Wand & Jones (1995) und Duong (2004). In der Doktorarbeit von Duong (2004) wurden auch statistische Verfahren ausführlich vorgestellt, mit denen man eine volle Bandbreitenmatrix nach dem MISE (Mean Integrated Squared Error) Kriterium bei Kerndichteschätzung bestimmen kann. Trotz des Fortschritts in der Berechnung des Glättungsparameters kann die wahre Dichte in manchen Fällen durch einen Kerndichteschätzer mit festem Glättungsparameter nicht gut widerspiegelt werden, insbesondere wenn die Daten einen langen Schwanz oder eine Multimodal-Struktur haben. In der multivariaten Datenanalyse ist diese Situation noch schlimmer, so dass ein Kerndichteschätzer mit festem Glättungsparameter in der Regel kein guter Schätzer ist, weil sie stark unter „Curse of Dimensionality“ leidet. Aus diesem Grund wurde in der Literatur (S. 202 von Scott (1992), S. 90 von Wand & Jones (1995)) erwähnt, dass ein Kerndichteschätzer mit festem Glättungsparameter für die Dichteschätzung in den Fällen nicht geeignet ist, wenn die Dimension der Daten größer 5 ist. Es sei denn, wenn der Umfang der Daten riesig ist. In der Arbeit von Scott & Wand (1991) wurden die äquivalenten Größen der Stichprobe in Bezug auf RCV (Root Coefficient of Variation), AMIAE (Asymptotic Mean Integrated Absolute Error) und RRMSE (Relative Root Mean Squared Error) über verschiedene Dimensionen (1-8) gezeigt, mit denen der gleiche Schätzungsfehler wie bei der Stichprobe mit Umfang 50 im univariaten Fall erreicht werden kann. Obwohl ein Kerndichteschätzer mit festem Glättungsparameter im hochdimensionalen Fall an sich kein guter Schätzer für die unbekannte wahre Dichte ist, bietet sich die Kerndichteschätzung in der multivariaten Datenanalyse als nützliches Werkzeug an, um die in den Daten versteckte Struktur aufzudecken. In dieser Arbeit wird die Anwendung der Kerndichteschätzung in der multivariaten Datenanalyse aus den folgenden Aspekten anhand simulierter und praktischer Daten erläutert:

1. Es gibt in der Regel keine statistische Methode, mit der man einen optimalen Glättungsparameter bei Kerndichteschätzung finden kann. „different useful information can be available at different levels of smoothing;“

2. Durch Einbezug vom gemischten Modell (inklusiv dem Kerndichteschätzer mit adaptivem Glättungsparameter) in die Datenanalyse kann die wahre Datenstruktur gut widerspiegelt werden. Man kann einen passenden Kerndichteschätzer mit festem Glättungsparameter als Pilot-Dichteschätzer für das Konstruieren eines gemischten Modells verwenden;

3. Das Dichteschätzer basierte Clusteringverfahren liefert ein gutes Resultat bei der Untersuchung der unbekannten Modal- bzw. Cluster-Struktur in den Daten;

4. Zum Zweck der Datenexploration gibt es im Prinzip kein optimales Prunen des

Dichteschätzer Basierten Dendrogramms. Mit einem Prunen des Dichteschätzer Basierten Dendrogramms erhält man nur eine gewisse Facette der Datenstruktur;

5. Man kann die Information aus Kerndichteschätzung bzw. darauf basierten statistischen Modellen auf eine geeignete Art und Weise mit in die Datenvisualisierung einbeziehen. Dies bietet eine Möglichkeit an, die Struktur der multivariaten Daten insbesondere der hochdimensionalen Daten zu visualisieren;

6. Wegen (1) und (4) spielt die grafische Darstellung in (5) beim Anwenden der Kerndichteschätzung in der explorativen Datenanalyse eine wichtige Rolle.

In diesem Einführungskapitel werden zwei univariate Datensätze benutzt, um die oben erwähnten in dieser Arbeit zu diskutierenden Probleme kurz vorzustellen. Da in dieser Arbeit ein Kerndichteschätzer mit sowohl festem als auch adaptivem Glättungsparameter verwendet wird, schreibt man einen Kerndichteschätzer mit festem bzw. adaptivem Glättungsparameter als Fest-Kerndichteschätzer bzw. Adaptiv-Kerndichteschätzer in den Stellen, in denen sie zusammen auftauchen. Ansonsten bezieht sich ein Kerndichteschätzer immer auf einen Kerndichteschätzer mit festem Glättungsparameter.

1.1 Auswahl des Glättungsparameters

Ein univariater Kerndichteschätzer $\hat{f}_k$ wird wie folgt definiert: Seien $X_1, ..., X_n \in R$ eine Zufallsstichprobe aus einer Wahrscheinlichkeitsverteilung mit Dichtefunktion f, dann

$$\hat{f}_k(x; h) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1.1}$$

wobei h der Glättungsparameter, K die Kernfunktion. Einen d-variaten Kerndichteschätzer definiert man entsprechend wie folgt: Seien $X_1, ..., X_n \in R^d$ eine Zufallsstichprobe aus einer Wahrscheinlichkeitsverteilung mit Dichtefunktion f, dann

$$\hat{f}_k(x; H) = (n|H|^{1/2})^{-1} \sum_{i=1}^{n} K(H^{-1/2}(x - X_i)) \tag{1.2}$$

wobei H eine $d \times d$ symmetrische und positiv definierte Bandbreitenmatrix, K die multivariate Kernfunktion. Zu bemerken ist, dass $H = h^2$ für $\hat{f}_k$ im univariaten Fall. Für eine ausführliche Beschreibung von Kerndichteschätzern verweisen wir auf die Arbeit von Silverman (1986), Scott (1992), Wand & Jones (1995), Simonoff (1996) und Duong (2004). Eine allgemeine Form $\hat{f}$ des d-variaten Kerndichteschätzers wurde in Scott & Szewczyk (2000) folgendermaßen definiert

$$\hat{f}(x, \theta) = \sum_{i=1}^{m} w_i f_i(x, \theta_i) \tag{1.3}$$

wobei $w_i > 0$ für $i = 1, ..., m$, $\sum_{i=1}^{m} w_i = 1$ und f_i eine beliebige Dichtefunktion mit Parameter θ_i.

Die Auswahl eines geeigneten Glättungsparameters spielt die entscheidende Rolle bei Kerndichteschätzung. Seit der Geburt des Kerndichteschätzers wurden verschiedene statistische Verfahren vorgeschlagen, um zu versuchen, einen optimalen Glättungsparameter in Bezug auf ein gewisses Kriterium zu bestimmen. Typische Bandbreitenselektoren sind LCV (Likelihood Cross Validation) von Habbema et al. (1974) und Duin (1976), LSCV (Least Squares Cross Validation) von Rudemo (1982) und Bowman (1984), BCV (Biased Cross Validation) von Scott & Terrell (1987), Direct-Plug-In Methode von Sheather & Jones (1991), SCV (Smoothed Cross Validation) von Hall et al. (1992). Für eine ausführliche Beschreibung der obigen Bandbreitenselektoren verweisen wir auf die Arbeit von Scott (1992), Wand & Jones (1995), Duong (2004), Li & Racine (2007). Vier Bandbreitenselektoren LCV, LSCV, BCV und Direct-Plug-In werden in Kapitel 2 anhand simulierter Daten vorgestellt und verglichen.

Im Folgenden wird anhand des **Hidalgo** Datensatzes veranschaulicht, wie die Auswahl des Glättungsparameters das Resultat der Kerndichteschätzung beeinflusst. Der **Hidalgo** Datensatz wurde von Izenman und Sommer (1988) in die statistische Literatur eingebracht und besteht aus 485 Messwerten von der Dicke der Hidalgo-Briefmarken, die im 19.Jahrhundert in Mexico gedruckt wurden. Das Ziel der Untersuchung ist, die Anzahl der Papierfabriken zu schätzen, die damals die Briefmarkenpapiere für die Hidalgo-Briefmarken herstellten. Im statistischen Sinne besteht das Ziel der Datenanalyse in der Festlegung der Anzahl der Modi in Dichtefunktion f. Da f unbekannt ist, wird hier die Modalstruktur von f anhand von $\hat{f}_k$ untersucht. Abbildung 1.1 zeigt $\hat{f}_k$ mit LCV, LSCV, BCV und Direct-Plug-In Bandbreiten. Ein paar Erklärungen dazu:

- Eine modifizierte Version der **Hidalgo** Daten (mit Hinzufügung von Störungstermen $\rho_i \sim U(-0,0005; 0,0005), i = 1, ..., 485$) wird hier verwendet, weil ansonsten man bei der LSCV Methode ein triviales Minimum ($h = 0$) nehmen müsste. Auf Details geht man in Abschnitt 2.2 ein;

- In Abbildung 1.1 verwendet man unterschiedliche y-Skala, um die Modi besser darzustellen.

In Abbildung 1.1 sind unterschiedliche Datenstrukturen unter verschiedenen Glättungsparametern zu erkennen, z.B., man sieht 7 Modi in $\hat{f}_k$ mit der Direct-Plug-In Bandbreite und nur 2 Modi in $\hat{f}_k$ mit der BCV Bandbreite. Es stellt sich nun die Frage, welche Modi von $\hat{f}_k$ auch tatsächlich in f existieren. Dieses Problem wurde in der Arbeit von Chaudhuri & Marron (1999) durch Nutzung eines grafischen Werkzeugs SiZer diskutiert. Die Grundidee von SiZer beruht auf der Scale Space Theorie (Lindeberg 1994) aus der Computer Vision, wobei man die Eigenschaften von f anhand von Dichteschätzern mit einer Reihe von Glättungsparametern untersucht. In der Tat ist es üblich, dass „different useful information can be available at different levels of smoothing.“ Mit einem

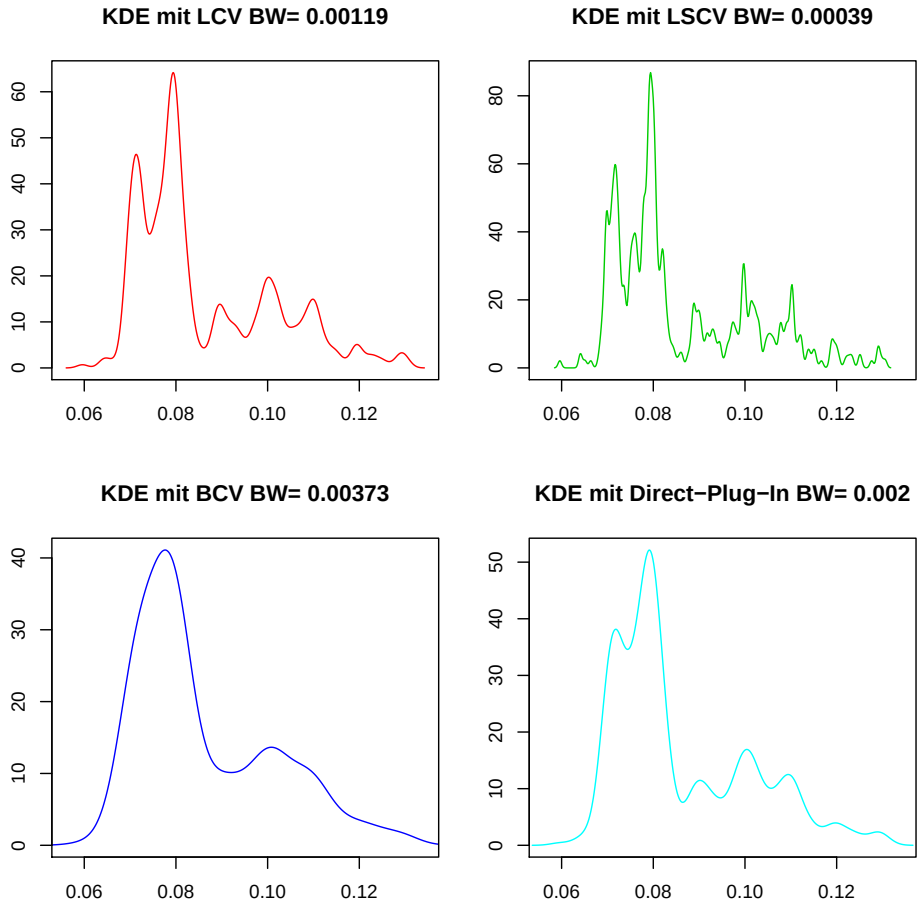

Abbildung 1.1: Kerndichteschätzer mit LCV (Grafik oben links), LSCV (Grafik oben rechts), BCV (Grafik unten links) und Direct-Plug-In (Grafik unten rechts) Bandbreiten bei den **Hidalgo** Daten

„optimalen“ Glättungsparameter bekommt man in der Regel nur eine Version der Daten, und

> ... all choices of the bandwidth h lead to useful density estimates. Large bandwidths provide a picture of the global structure in the unknown density, ... small bandwidths, ..., reveal local structure which may or may not be present in the true density.
>
> Scott (1992)

Abbildung 1.2 zeigt den SiZer Plot von Chaudhuri & Marron (1999) für die **Hidalgo** Daten, wobei das blaue bzw. rote Gebiet für die Stellen steht, an denen $\hat{f}_h'$ signifikant steigend bzw. fallend ist ($\hat{f}_h'$ ist die erste Ableitung von $\hat{f}_k(x,h)$). Man markiert im SiZer Plot die LCV bzw. Direct-Plug-In Bandbreite mit einer weißen bzw. grünen Linie, um den SiZer Plot mit den entsprechenden Kerndichteschätzern in Abbildung 1.1 zu vergleichen. Im SiZer Plot werden diejenige Modi, deren Nullstellen von $\hat{f}_h'$ zwischen einem blauen und einem roten Bereich liegen, als signifikant bezeichnet (Details vgl. Chaudhuri & Marron (1999)). Aus dem SiZer Plot in Abbildung 1.2 ist der folgende

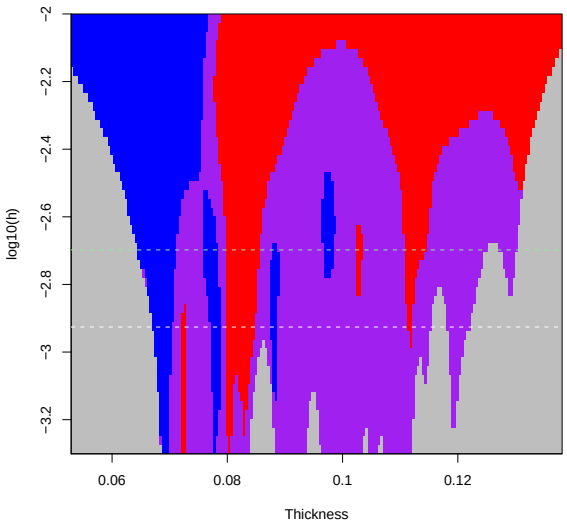

Abbildung 1.2: SiZer Plot von Chaudhuri & Marron (1999) für die **Hidalgo** Daten

Schluss zu ziehen:

- Die Modi an $x = \{0,072; 0;08; 0,10\}$ sind tatsächlich vorhanden, weil sie zwischen einem blauen und einem roten Bereich liegen;
- Die Modi an $x = \{0,09; 0,11\}$ sind zweifelhaft;

- Die Modi an $x = \{0,12; 0,13\}$ sind nur ein Artefakt aus der Stichprobenerhebung.

Im multivariaten Fall ist das Problem schwierig zu lösen, weil die multivariate Datenstruktur nicht direkt zu veranschaulichen ist. Darauf geht man in Kapitel 3 und 4 ein.

1.2 Kerndichteschätzer und Gemischtes Modell

Die allgemeine Form des Kerndichteschätzers in (1.3) stellt eigentlich ein gemischtes Modell dar. Wenn man f_i in (1.3) als Normaldichte annimmt, dann bekommt man

$$\hat{f}(x, \theta) = \sum_{i=1}^{m} w_i \phi_i(x, \theta_i) \tag{1.4}$$

mit Normaldichte ϕ_i und $\theta = (\mu_i, \Sigma_i)$, wobei μ_i der Erwartungswert und Σ_i die Varianz-Kovarianz-Matrix der entsprechenden Normalverteilung. Typische Varianten von dem Modell in (1.4) sind Adaptiv-Kerndichteschätzer von Sain (1999, 2002) und Gemischtes Modell aus dem Model Based Clustering (Fraley & Raftery (2002)). Ein für die Daten gut geeignetes gemischtes Modell erhält man im Prinzip entweder durch den EM-Algorithmus (Celeux & Govaert (1992, 1995), Fraley & Raftery (2002)) oder durch Fusion der Komponenten eines Pilot-Fest-Kerndichteschätzers (Scott & Szewczyk (2000), Sain (2002)). Im Folgenden wird dieses Problem anhand eines praktischen Datensatzes kurz vorgestellt. In Kapitel 3 geht man darauf weiter ein.

Beispiel 1.2.1 Daten aus dem Sportzentrum der Universität Augsburg

Das Tempo von 3646 Bewegungen beim Fußballtraining wurde gemessen. Die 3646 Messwerte teilt man in vier Kategorien (*Stehen*, *Gehen*, *Traben* und *High Intensiv*) auf. Man nimmt an, dass die Messwerte aus einer unbekannten Wahrscheinlichkeitsverteilung mit Dichtefunktion f kommen. Man wollte wissen, ob einer gemischten Normalverteilung die 3646 Messwerte unterliegen.

Abbildung 1.3 zeigt die Kerndichteschätzer für f mit 7 verschiedenen Bandbreiten, wobei die 7 Kerndichteschätzer unterschiedlich eingefärbt sind. In Abbildung 1.3 sieht man, dass

- eine 4-Modi Struktur in den Kerndichteschätzern zu erkennen ist;
- es Probleme am linken Rand gibt, weil das Tempo der Bewegung beim Fußballtraining auf keinen Fall kleiner Null sein kann.

Ein sinnvoller Lösungsvorschlag für das Randproblem bei Kerndichteschätzung ist durch Nutzung einer Randkernfunktion. Man kann z.B. die folgende Floating-Randkernfunktion von Scott (1992)

$$K_c(t) = \frac{3}{4}[(c+1) - \frac{5}{4}(1+2c)(t-c)^2](t-(c+2))^2 I_{[c,c+2]}(t) \tag{1.5}$$

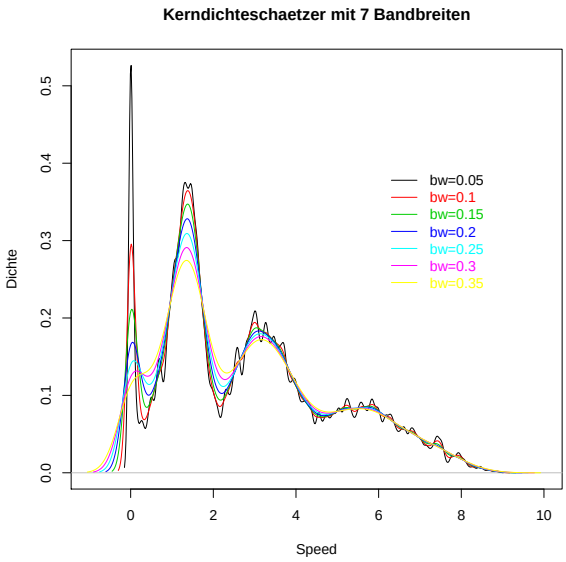

Abbildung 1.3: Kerndichteschätzer mit 7 verschiedenen Bandbreiten für f in Beispiel 1.2.1

benutzen, um f im linken Randbereich besser zu schätzen. Abbildung 1.4 zeigt den entsprechenden Kerndichteschätzer mit Randkernfunktion $K_c(t)$ für $x \in [0; h)$ und der Biweight Kernfunktion für $x \in [h; 1,05 \cdot max(x_1, ..., x_n)]$ mit Glättungsparameter $h = 0,5$. In Abbildung 1.4 ist keine Form einer gemischten Normalverteilung zu erkennen, deren Hauptgrund darin liegt, dass es 213 Nulls in den Messungen aus Kategorie *Stehen* gibt und die Daten am linken Rand keiner Normalverteilung folgen.

Im Folgenden zeigt man, dass eine gemischte Normalverteilung den Messwerten aus Kategorien *Gehen*, *Traben* und *High Intensiv* gut anpasst. Abbildung 1.5 zeigt die Kerndichteschätzer für f_{os} mit 7 verschiedenen Bandbreiten, wobei f_{os} für die Dichtefunktion der Verteilung der Daten aus Kategorien *Gehen*, *Traben* und *High Intensiv* steht. In Abbildung 1.5 ist eine 3-Modi Struktur klar zu erkennen. Um zu zeigen, dass die Daten aus einer gemischten Normalverteilung mit 3 Komponenten kommen, geht man wie folgt vor:

- Durch Nutzung vom SiZer Plot in Abbildung 1.6 zeigt man, dass die 3 Modi tatsächlich in f_{os} sind. Man sieht in Abbildung 1.6, dass zwei Modi (links und mittel) zwischen einem blauen und einem roten Bereich liegen, was impliziert, dass die zwei Modi signifikant sind. Aus der Tatsache, dass der Bereich zwischen dem Modus in der Mitte und dem rechten Modus relativ groß ist und die Kurve auf der rechten Seite des rechten Modus signifikant fallend ist, folgt dass der rechte Modus auch tatsächlich da ist;

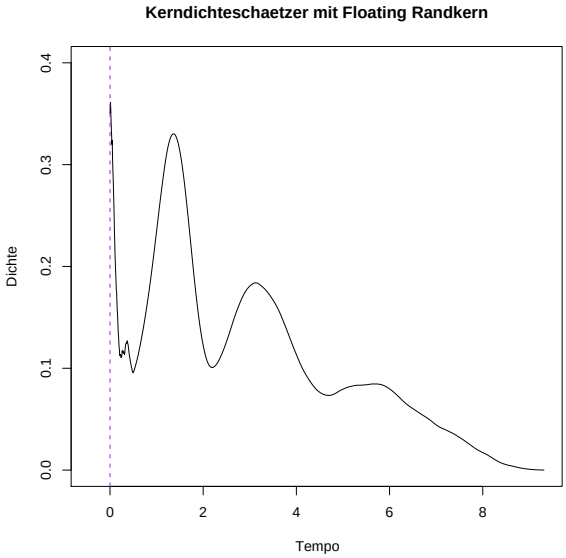

Abbildung 1.4: Kerndichteschätzer mit Randkernfunktion $K_c(t)$ für $x \in [0, h)$ und der Biweight Kernfunktion für $x \in [h; 1,05 \cdot max(x_1, ..., x_n)]$ mit Glättungsparameter $h = 0,5$ für die Daten in Beispiel 1.2.1

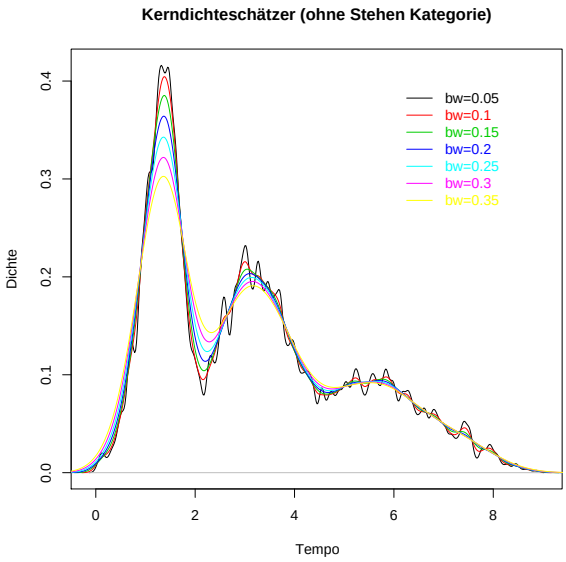

Abbildung 1.5: Kerndichteschätzer mit 7 verschiedenen Bandbreiten für f_{os} in Beispiel 1.2.1

- Man berechnet den Dichteschätzer $\hat{f}_{os}^{m}$ für f_{os} mittels des Model Based Clustering Verfahrens. Das beste Modell ist das gemischte Modell mit 3 Komponenten, deren Mittelwerte an $1,31$, $3,13$ und $5,56$ liegen. Der Unterschied der BIC-Werte zwischen dem besten Modell (-12365,59 bei **V** Modell mit 3 Komponenten) und zweitbesten Modell (-12392,00 bei **V** Modell mit 4 Komponenten) beträgt $26,41$, was auf einen signifikanten Unterschied zwischen den zwei Modellen hinweist (Fraley & Raftery (1999));

- Man vergleicht das Clustering aus dem Model Based Clustering mit den vorgegebenen Kategorien und stellt das Resultat im Fluctuationsdiagramm in Abbildung 1.7 dar. Man sieht in Abbildung 1.7, dass das 3-Cluster-Modell aus dem Model Based Clustering die vorgegebene Klassifizierung der Daten in Kategorien *Gehen*, *Traben* und *High Intensiv* gut widerspiegelt;

- Man konstruiert das gemischte Modell $\hat{f}_{os}^{k}$ für f_{os} mit dem modifizierten SEM-Algorithmus (vgl. Abs. 3.1), wobei der Kerndichteschätzer mit $h = 0,3$ als Pilot-Kerndichteschätzer verwendet wird. In Abbildung 1.8 stellt man die Kerndichteschätzer mit $h = \{0,1; 0,2; 0,3\}$ und zwei gemischte Modelle (**V** Modell mit 3 Komponenten aus dem Model Based Clustering und das gemischte Modell aus dem modifizierten SEM-Algorithmus) grafisch dar. Die Grafik in Abbildung 1.8 liefert einen starken Hinweis darauf, dass die Daten einer gemischten Normalverteilung mit 3 Komponenten unterliegen.

In unserem Beispiel werden die Daten anhand verschiedener Modelle untersucht, um keine nützliche Information zu verlieren. Die Vorteile dieser Modelle in der explorativen Datenanalyse können wie folgt beschrieben werden:

1. Ein Kerndichteschätzer ist in der Regel ein „Overparametrized“ Dichteschätzer, in dem mehr lokale Eigenschaften der Daten zu erkennen sind;

2. Das Modell aus dem Model Based Clustering liefert ein ziemlich gutes Resultat, wenn die Modellannahme stimmt;

3. Das Modell aus dem SEM Algorithmus stellt einen Kompromiss dar, das in der Situation verwendet werden kann, wenn es keine sichere Modellannahme gibt und ein „Overparametrized“ Modell scheitert.

1.3 Dichteschätzung und Clustering

Eine der wichtigsten Anwendungen der Dichteschätzung in der explorativen Datenanalyse liegt in der Identifizierung der Modalstruktur (Mode Hunting) und dem darauf basierten Clustering. Grob kann man die Dichteschätzer basierten Clusteringmethoden

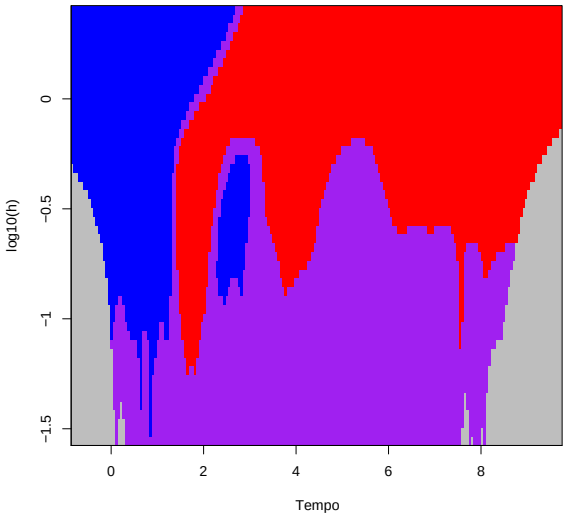

Abbildung 1.6: SiZer Plot für die Daten in Beispiel 1.2.1 ohne Kategorie *stehen*

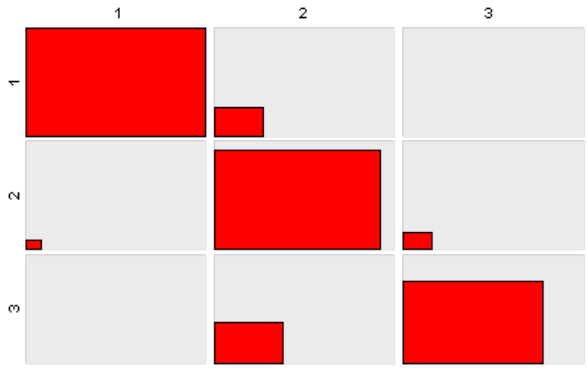

Abbildung 1.7: Vergleich des 3-Cluster-Modells aus dem Model Based Clustering und der vorgegebenen drei Kategorien *Gehen*, *Traben* und *High Intensiv* in Beispiel 1.2.1 im Fluctuationsdiagramm

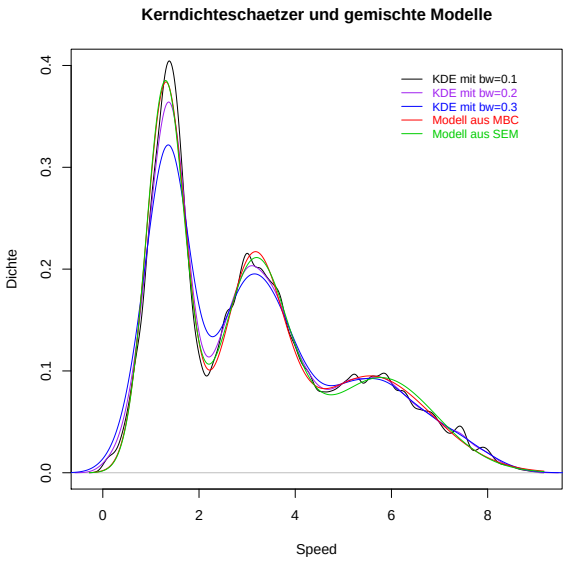

Abbildung 1.8: Kerndichteschätzer mit $h = \{0,1; 0,2; 0,3\}$, Gemischtes Modell mit 3 Komponenten aus dem Model Based Clustering, Gemischtes Modell aus dem SEM Algorithmus mit dem Kerndichteschätzer mit $h = 0,3$ als Pilot-Dichteschätzer in Beispiel 1.2.1

in parametrisches und nichtparametrisches Clustering aufteilen. Eine typische parametrische Methode ist das Model Based Clustering von Fraley & Raftery (2002). Beim nichtparametrischen Dichteschätzer basierten Clustering wird angenommen, dass die Daten einer gewissen Wahrscheinlichkeitsverteilung mit Dichtefunktion f unterliegen und zu den Domänen der Modi in f gehören. Das nichtparametrische Dichteschätzer basierte Clustering wurde zuerst in Wishart (1969) untersucht und dann von Hartigan (1975) erweitert, indem er den Begriff „High Density Clusters“ in die Dichteschätzer basierte Clusteranalyse einführte, die als zusammenhängende Komponenten in der λ-Niveaumenge

$$L(\lambda, f) = \{x | f(x) > \lambda\}$$

definiert worden sind. Die Grundidee von Hartigan liegt darin, dass falls die Daten eine Clusterstruktur haben, dann sollte diese Struktur in den Hohe-Dichte-Regionen auffällig und gut zu identifizieren sein. Da f in der Praxis meist unbekannt ist, schätzt man f durch einen Dichteschätzer $\hat{f}$ und zieht die Modalstruktur von $\hat{f}$ in Betracht, um die Daten der geschätzten Modalstruktur entsprechend in Clustern aufzuteilen. Im Folgenden wird das nichtparametrische Dichteschätzer basierte Clustering anhand der Daten in Beispiel 1.2.1 (ohne Kategorie *Stehen*) kurz vorgestellt.

Man geht wie folgt vor:

- Man schätzt f durch Kerndichteschätzer $\hat{f}_{os}^{k}$ mit $h = 0,3$;

- Man zeigt die geschätzte Clusterstruktur der Daten den High Density Clustern in $\hat{f}_{os}^k$ entsprechend in Abbildung 1.9. Man sieht in Abbildung 1.9, dass es zwei High Density Cluster in $L(0,086; \hat{f}_{os}^k)$ gibt und die Submenge $L(0,135; \hat{f}_{os}^k)$ des linken Clusters in zwei High Density Subclustern auf $\lambda = 0,135$ gesplittet wird;
- In Abbildung 1.9 sind 3 Modi zu erkennen. Man ordnet die Daten mit dem Hill-Climbing Algorithmus von Hinneburg & Gabriel (2007) zu den 3 Modi zu und stellt diese Zuordnung der Daten in Abbildung 1.10 dar;
- Man vergleicht das Resultat aus dem obigen nichtparametrischen Clustering mit den vorgegebenen Kategorien und stellt die Konfusionsmatrix tabellarisch in Tabelle 1.1 und grafisch im Fluctuationsdiagramm in Abbildung 1.11 dar.

Mode	Cluster rechts	Cluster mittel	Cluster links
Gehen	1213	5	0
High Intensiv	136	1118	59
Traben	0	79	676

Tabelle 1.1: Konfusionsmatrix von dem Resultat aus dem nichtparametrischen Clustering und den vorgegebenen Kategorien *Gehen*, *Traben* und *High Intensiv* in Beispiel 1.2.1

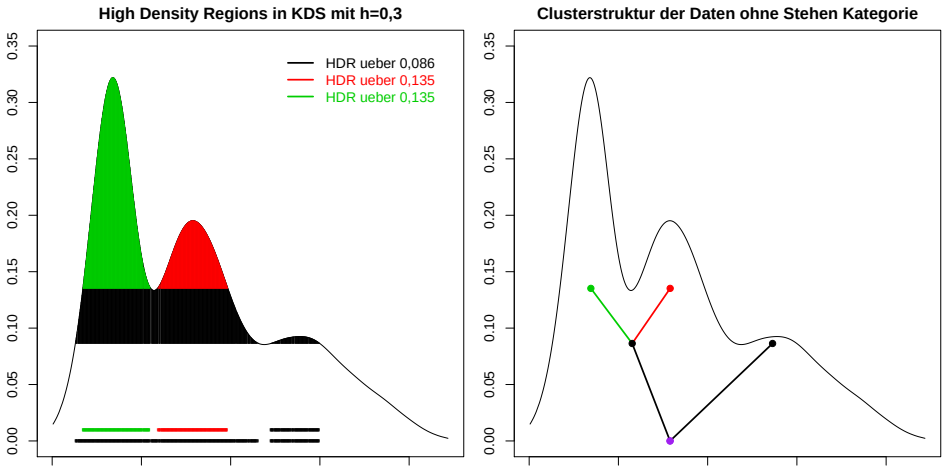

Abbildung 1.9: Geschätzte Clusterstruktur der Daten (ohne Kategorie *Stehen*) in Beispiel 1.2.1

In Abbildung 1.9-1.11 sieht man, dass das nichtparametrische Dichteschätzer basierte Clusteringverfahren ein sinnvolles Resultat liefert im Vergleich zu den vorgegebenen Kategorien. Damit hat man auch die Möglichkeit, die Clusterstruktur der Daten zu

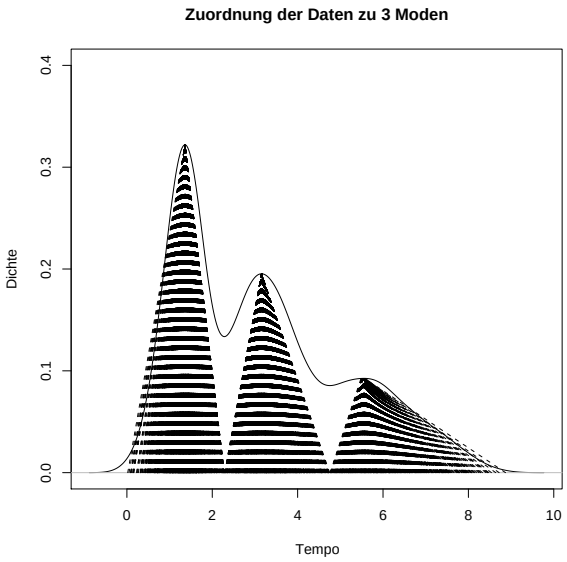

Abbildung 1.10: Zuordnung der Daten zu den 3 Modi in $\hat{f}_{os}^{k}$ für die Daten aus Kategorien *Gehen*, *Traben* und *High Intensiv* in Beispiel 1.2.1

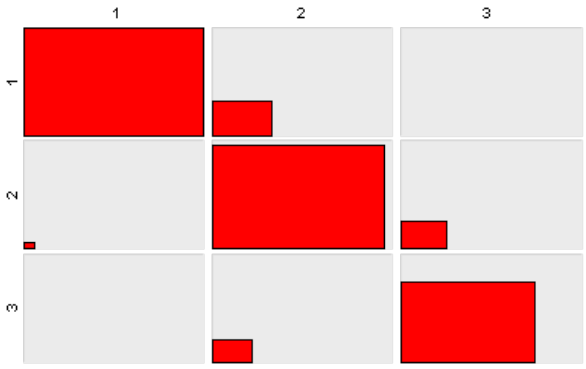

Abbildung 1.11: Vergleich des Resultats aus dem nichtparametrischen Clustering mit den vorgegebenen 3 Kategorien *Gehen*, *Traben* und *High Intensiv* im Fluctuationsdiagramm

schätzen und zu visualisieren (vgl. Abbildung 1.9). Ein Nachteil des nichtparametrischen Dichteschätzer basierten Clusterings liegt daran, dass dessen Resultat stark von dem Dichteschätzer abhängt.

Das Dichteschätzer basierte Clusteringverfahren bietet sich in der explorativen Datenanalyse als nützliches Werkzeug an insbesondere im multivariaten Fall. In Kapitel 3 geht man darauf ausführlich ein.

1.4 Struktur der vorliegenden Arbeit

In der vorliegenden Arbeit wird die Anwendung der nichtparametrischen Dichteschätzung in der explorativen Datenanalyse anhand simulierter und praktischer Daten erläutert. Wie oben erwähnt konzentriert sich die Arbeit auf die Kerndichteschätzung mit Gaussian Kernfunktion als Default. In Kapitel 2 werden vier Bandbreitenselektoren (LCV, LSCV, BCV und Direct-Plug-In) im uni- und multivariaten Fall anhand simulierter Daten verglichen, um das Verständnis der Rolle des Glättungsparameters bei Kerndichteschätzung zu vermitteln. Zwei übliche Probleme in der praktischen Anwendung von LSCV und BCV Methoden werden auch in Kapitel 2 diskutiert, die sind LSCV bei Datensätzen mit Bindungen, und wie man beim Anwenden der BCV Methode einen passenden Glättungsparameter aus mehreren Kandidaten auswählt. In Kapitel 3 wird der Zusammenhang von Kerndichteschätzung mit gemischtem Modell und mit dem Dichteschätzer basierten Clustering anhand zwei praktischer Datensätze diskutiert. In der explorativen Datenanalyse besteht ein enger Zusammenhang zwischen dem Kerndichteschätzer, gemischtem Modell und Dichteschätzer basiertem Clustering. Gute Beispiele findet man in Scott & Szewczyk (2000), Sain (2002), Fraley & Raftery (1999,2002), Stuetzle (2003), Stuetzle et al. (2007) usw.. Die Datenvisualisierung bietet sich als nützliches Werkzeug in der explorativen Datenanalyse an, deren Hauptaufgabe darin besteht, die Eigenschaften der Daten wie Muster, Cluster, Korrelation usw. aufzudecken. In Kapitel 4 werden verschiedene Visualisierungstechniken in der multivariaten Datenanalyse vorgestellt. Der Schwerpunkt von Kapitel 4 liegt an der grafischen Darstellung der Information aus dem Dichteschätzer basierten Modell. Die statistische Software **R** wird in dieser Arbeit für die Unterstützung der Argumente intensiv eingesetzt. In den letzten Jahren wurde eine Vielzahl von **R** Paketen entwickelt, auch im Bereich von Kerndichteschätzung. Kapitel 5 gibt einen Überblick über die relevanten **R** Pakete für Kerndichteschätzung. Schließlich wird alles zusammengefasst und einen Ausblick gegeben.

2 Kerndichteschätzung und Glättungsparameter

Die Auswahl des Glättungsparameters hat einen großen Einfluss auf Kerndichteschätzung. In den letzten Jahren wurde eine Vielzahl von Methoden vorgeschlagen, um einen sowohl theoretisch als auch praktisch zur Kerndichteschätzung gut passenden Glättungsparameter zu bestimmen. In diesem Sinne ist die Auswahl einer geeigneten Bandbreite eigentlich das Problem der Auswahl eines geeigneten Bandbreitenselektors. In der Arbeit von Loader (1999) wurden die Vor- und Nachteile der klassischen Methoden (LCV, LSCV/AIC) und Plug-In Methoden (BCV, Direct-Plug-In) im univariaten Fall gut untersucht. Während die damalige Stimme mehr für die Plug-In Methoden war, sagte Loader, dass die Plug-In Methoden stark vom Pilot-Glättungsparameter abhängig sind und damit man wichtige Eigenschaft der Daten verpassen kann, und die große Variabilität und mögliche Unterglättung bei klassischen Methoden eigentlich die Unsicherheit beim Auswählen des Glättungsparameters widerspiegeln. In der Praxis sind klassische Methoden aber in manchen Situationen schlecht anwendbar, z.B., LCV beim Datensatz mit dickem Tail und Ausreißern, LSCV beim Datensatz mit diskretisierter Variable.

In der Literatur werden zwei Gütekriterien oft für den Schätzer des einem gewissen Kriterium z.B. MISE entsprechenden optimalen Glättungsparameters benutzt: Konvergenzrate gegen den wahren optimalen Glättungsparameter und Variabilität des geschätzten Glättungsparameters beim Datensatz mit endlichem Umfang. In diesem Kapitel werden vier Bandbreitenselektoren (zwei Klassische: LCV und LSCV, zwei Plug-In: BCV und Direct-Plug-In) im uni- und multivariaten Fall anhand simulierter Daten verglichen, um zu zeigen, wie gut man den MISE-optimalen Glättungsparameter mit den vier Bandbreitenselektoren aus einer Stichprobe mit endlichem Umfang schätzen kann. Für interessante Diskussion über die Konvergenzrate von Plug-In und Cross-Validation Bandbreiten gegen den MISE-optimalen Glättungsparameter verweisen wir auf die Arbeit von Duong & Hazelton (2003) sowie Duong (2004). Das Ziel dieser Untersuchung besteht darin, die Rolle des Glättungsparameters bei Kerndichteschätzung besser zu verstehen.

Das vorliegende Kapitel gliedert sich in 4 Teile. Abschnitt 2.1 gibt einen Überblick über die vier Bandbreitenselektoren und die in diesem Kapitel verwendeten Rechenmethoden an. In Abschnitt 2.2 werden die vier Methoden anhand der 1D (**Hidalgo**), 2D (**geyser**) und 5D (simulierten) Daten verglichen. In Abschnitt 2.3 werden zwei Probleme bei der Anwendung der CV-Methoden (LSCV und BCV) in der Datenanalyse diskutiert. Schließlich fasst man alles in Abschnitt 2.4 zusammen.

2.1 Überblick über die vier Methoden

Wie in (1.2) steht wird ein d-variater Kerndichteschätzer wie folgt definiert: Seien $X_1, X_2, ..., X_n \in R^d$ eine Zufallsstichprobe aus einer Wahrscheinlichkeitsverteilung mit unbekannter Dichtefunktion f, dann ist $\hat{f}(x; H)$ mit

$$\hat{f}(x; H) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i) \tag{2.1}$$

ein d-variater Kerndichteschätzer für f, wobei $x = (x_1, x_2, ..., x_d)^T$, $X_i = (X_{i1}, X_{i2}, ..., X_{id})^T$ für $i = 1, ..., n$, H eine $d \times d$ symmetrische und positiv definierte Bandbreitenmatrix, $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$, wobei K die Kernfunktion. Zur Einfachheit der grafischen Darstellung beschränken wir uns in diesem Kapitel auf Kerndichteschätzer $\hat{f}(x; h)$ mit Produktkern K und diagonaler Bandbreitenmatrix h. Der Kerndichteschätzer $\hat{f}(x; h)$ sieht entsprechend wie folgt aus:

$$\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \tag{2.2}$$

wobei $h = (h_1, h_2, ..., h_d)^T$ und K_h der skalierte Produktkern mit

$$K_h(x - X_i) = \frac{1}{\left(\prod_{l=1}^{d} h_l\right)} K\left(\frac{x_1 - X_{i1}}{h_1}\right) K\left(\frac{x_2 - X_{i2}}{h_2}\right) ... K\left(\frac{x_d - X_{id}}{h_d}\right) \tag{2.3}$$

Das MISE Gütekriterium für $\hat{f}(x; h)$ bei Kerndichteschätzung definiert man wie folgt:

$$MISE\hat{f}(x; h) = E[ISE\hat{f}(x; h)] = E \int_{R^d} [\hat{f}(x; h) - f(x)]^2 dx \tag{2.4}$$

wobei

$$ISE\hat{f}(x; h) = \int_{R^d} [\hat{f}(x; h) - f(x)]^2 dx.$$

Ein MISE-optimaler Glättungsparameter h_{MISE} ist dann $h_{MISE} = argmin_{h \in R^d} MISE\hat{f}(x; h)$. Der LSCV-Glättungsparameter $\hat{h}_{lscv}$ ist ein unbiased Schätzer für h_{MISE} in dem Sinne, dass

$$\hat{h}_{lscv} = argmin_{h \in R^d} MISE_{lscv}\hat{f}(x; h)$$

mit

$$MISE_{lscv}\hat{f}(x; h) = MISE\hat{f}(x; h) - \int f(x)^2 dx$$

weil

$$MISE\hat{f}(x;h) = E\int \hat{f}(x;h)^2 dx - 2E\int \hat{f}(x;h)f(x)dx + \int f(x)^2 dx \tag{2.5}$$

und f die wahre Dichtefunktion und nicht von h abhängig ist.

Da das Gütekriterium $MISE$ im Allgemeinen auf eine komplizierte Art und Weise von h abhängt (Wand & Jones (1995)), wird in der Praxis oft eine asymptotische Form $AMISE$ (Asymptotic Mean Integrated Squared Error) von $MISE$ mit

$$AMISE\hat{f}(x;h) = \frac{R(K)}{n\prod_{l=1}^{d} h_l} + \frac{1}{4}\mu_2(K)^2 (h^2)^T \Psi_D (h^2) \tag{2.6}$$

wobei $R(K) = \int K(z)^2 dz$, $\mu_2(K)I = \int zz^T K(z)dz$, $h = (h_1, ..., h_d)^T$, Ψ_D eine $d \times d$ Matrix, deren (i,j)-tes Element gleich $\psi_{2e_i+2e_j}$ ist, wobei e_i ein d-Tupel mit 1 an der i-ten Position und 0 sonst (Details vgl. S. 7 von Duong (2004)), für die Bestimmung des Glättungsparameters bei Kerndichteschätzung verwendet aber unter gewissen Annahmen (Wand & Jones (1995)). Bei Plug-In Methoden wird der AMISE-optimale Glättungsparameter h_{AMISE} mit $h_{AMISE} = argmin_{h\in R^d} AMISE\hat{f}(x;h)$ geschätzt.

Im Unterschied dazu hat die LCV Methode einen anderen mathematischen Hintergrund: Der Kerndichteschätzer $\hat{f}$ wird als eine parametrische Familie von Dichten mit Parameter h betrachtet, gegeben $X_1, ..., X_n$. Bei der LCV Methode kann man zeigen, dass die KLI (Kullback-Leibler Information) beim Maximieren der LCV Score-Funktion minimiert wird aber unter starken Annahmen an f (Hall (1987)). Im Folgenden werden die vier Methoden und die Hauptkritik daran kurz vorgestellt.

2.1.1 LCV

Das LCV Kriterium wurde von Habbema, Hermans und Van den Broek (1974) und Duin (1976) vorgeschlagen. Die Score-Funktion bei der LCV Methode definiert man wie folgt:

$$LCV(h) = \frac{1}{n}\sum_{i=1}^{n} log\hat{f}_{-i}(X_i, h) \tag{2.7}$$

wobei

$$\hat{f}_{-i}(X_i, h) = \frac{1}{(n-1)\left(\prod_{l=1}^{d} h_l\right)} \sum_{j=1, j\neq i}^{n} \prod_{l=1}^{d} K\left(\frac{X_{il} - X_{jl}}{h_l}\right) \tag{2.8}$$

Der Glättungsparameter wird bei der LCV Methode durch $\hat{h}_{lcv}$ mit $\hat{h}_{lcv} = argmax_{h\in R^d}(LCV(h))$ geschätzt.

Als Rechenmethode verwendet man in dieser Arbeit eine quasi-Newton Methode (BFGS von Broyden, Fletcher, Goldfarb und Shanno (1970)), um die Score-Funktion $LCV(h)$

in folgender Form zu optimieren:

$$LCV(h) = \log\left(\frac{1}{[(n-1)\prod_{l=1}^{d} h_l(\sqrt{2\pi})^d]^n}\right) + \sum_{i=1}^{n}\log\left(\sum_{j=1}^{n}\left(\prod_{l=1}^{d}\exp\left(-\frac{(X_{jl}-X_{il})^2}{2h_l^2}\right)-1\right)\right) \tag{2.9}$$

Die Hauptkritik an der LCV Methode liegt an den folgenden Punkten (Silverman (1986)):

1. Die LCV Methode ist sensitiv gegenüber Ausreißern;
2. Die Methode versagt bei Daten mit dickem Schwanz in dem Fall, wenn eine Kernfunktion mit dünnem Schwanz verwendet wird;
3. Es wurde gezeigt, dass der LCV Schätzer wegen des Schwanzes in den Daten nicht konsistent ist;
4. Die LCV Methode hat Probleme bei Daten mit diskretisierten Variablen.

2.1.2 LSCV

Die LSCV Methode wurde zuerst von Rudemo (1982) und Bowman (1984) vorgeschlagen. Die Zielfunktion bei der LSCV Methode wird wie folgt definiert (Li & Racine (2007)):

$$LSCV(h) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_h * K_h)(X_i - X_j) - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}K_h(X_i - X_j) \tag{2.10}$$

wobei das Zeichen * für die Faltung steht. Der Glättungsparameter wird bei der LSCV Methode durch $\hat{h}_{lscv}$ mit $\hat{h}_{lscv} = argmin_{h\in R^d}(LSCV(h))$ geschätzt.

Als Rechenmethode verwendet man in dieser Arbeit eine modifizierte quasi-Newton Methode (L-BFGS-B von Byrd et. al. (1995)) für univariate Daten und die BFGS Methode für multivariate Daten, um die Zielfunktion bei der LSCV Methode in folgender Form zu optimieren:

$$LSCV(k) = \frac{\prod_{l=1}^{d} k_l^{1/2}}{n(\sqrt{2\pi})^d}\left(\frac{1}{2^{d/2}} + \frac{n-1}{2^{d+3}} + \frac{1}{n}\sum_{j<i}^{i=2,\ldots,n}\left[-4\left(\prod_{l=1}^{d} Z_{lij}^{k_l} - \frac{1}{4\cdot 2^{d/2}}\right)^2\right]\right) \tag{2.11}$$

wobei $Z_{lij} = \exp^{-(X_{il}-X_{jl})^2/4}$ und $k = (k_1, \ldots k_d)^T$ mit $k_l = 1/h_l^2$ für $l \in (1, \ldots, d)$.

Die Kritikpunkte an der LSCV Methode und mögliche Lösungsvorschläge werden wie folgt aufgelistet:

1. $\hat{h}_{lscv}$ hat eine große Variabilität;

2. In manchen Fällen kommen im gesuchten Bereich mehrere lokale Minima von $LSCV(h)$ vor. In der explorativen Datenanalyse soll man in solcher Situation mit der Auswahl des Glättungsparameters vorsichtig umgehen. Ein möglicher Lösungsvorschlag dazu ist, dass man das größte lokale Minimum auswählt;

3. $\hat{h}_{lscv}$ konvergiert im univariaten Fall langsam gegen h_{MISE}. In der Arbeit von Duong & Hazelton (2005) wurde aber gezeigt, dass ein allgemeiner LSCV Bandbreitenschätzer $\hat{H}_{lscv}$ im multivariaten Fall mit der Konvergenzrate $n^{-min(d,4)/(2d+8)}$ gegen H_{MISE} ein besseres Verhalten zeigt;

4. Ein anderes bekanntes Problem der LSCV Methode ist, dass die Zielfunktion für kleinen Glättungsparameter hochsensitiv gegenüber „very fine small-scale effects“ ist (Silverman (1986)). Es wurde gezeigt, dass wenn die Daten diskretisierte Variable(n) enthalten, existiert in manchen Fällen kein lokales Minimum im Intervall von Null bis h_{OS} (Oversmoothed Bandbreite) (vgl. S. 165-166 in Scott (1992)). Auf dieses Problem geht man in Abschnitt 2.3 ausführlich ein. Ein nützlicher Trick bei Anwendung der LSCV Methode in dieser Situation ist, dass man simulierte Störungsterme zu diskretisierten Variablen hinzufügt, z.B., simulierte gleichverteilte Terme (Zychaluk & Patil (2006)).

2.1.3 BCV

Die BCV Methode wurde zuerst von Scott & Terell 1987 vorgeschlagen. Die Zielfunktion bei der BCV Methode definiert man wie folgt (Wand & Jones (1995)):

$$BCV(h) = \frac{R(K)}{n\prod_{l=1}^{d} h_l} + \frac{1}{4}\mu_2(K)^2 (h^2)^T \Psi_D(h^2) \tag{2.12}$$

Zum besseren Verständnis dieser Darstellung wird im Folgenden die $BCV(h)$ im bivariaten Fall als Beispiel gezeigt. Die BCV Zielfunktion im bivariaten Fall kann wie folgt geschrieben werden (Sain et al. (1994)):

$$\begin{aligned} BCV(h) = \frac{R(K)}{nh_1h_2} + \frac{1}{4}\sigma_K^4 [h_1^4 \int\int (f^{(2,0)}(x_1,x_2))^2 dx_1 dx_2 \\ + h_2^4 \int\int (f^{(0,2)}(x_1,x_2))^2 dx_1 dx_2 + 2h_1^2 h_2^2 \int\int f^{(2,0)}(x_1,x_2) f^{(0,2)}(x_1,x_2) dx_1 dx_2] \end{aligned} \tag{2.13}$$

wobei $h = (h_1, h_2)^T$, $\sigma_K^2 = \int u^2 K du$ (für K in (2.3)), $f^{(2,0)}(x_1,x_2) = \frac{\partial^2}{\partial x_1 \partial x_1} f(x)$ und $f^{(0,2)}(x_1,x_2) = \frac{\partial^2}{\partial x_2 \partial x_2} f(x)$. Aus partieller Integration kann gezeigt werden, dass

$$\int\int (f^{(2,0)}(x_1,x_2))^2 dx_1 dx_2 = E\left[\frac{\partial^4 f(x_1,x_2)}{\partial x_1^4}\right] \tag{2.14}$$

$$\int\int (f^{(0,2)}(x_1,x_2))^2 dx_1 dx_2 = E\left[\frac{\partial^4 f(x_1,x_2)}{\partial x_2^4}\right] \tag{2.15}$$

und

$$\int\int f^{(2,0)}(x_1,x_2) f^{(0,2)}(x_1,x_2) dx_1 dx_2 = E\left[\frac{\partial^4 f(x_1,x_2)}{\partial x_1^2 \partial x_2^2}\right] \tag{2.16}$$

Es gibt zwei Varianten von $BCV(h)$, deren Unterschied an der Schätzung von Ψ_D also eigentlich von ψ_r liegt. In dieser Arbeit bezeichnet man die zwei Varianten von $BCV(h)$ mit $BCV1(h)$ und $BCV2(h)$ und die zwei entsprechenden Schätzer von ψ_r mit $\check{\psi}_r$ und $\tilde{\psi}_r$, die üblicherweise wie folgt definiert werden (Duong & Hazelton 2005):

$$\check{\psi}_r(h) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}(K_h^{(r)} * K_h)(X_i - X_j) \tag{2.17}$$

und

$$\tilde{\psi}_r(h) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}K_h^{(r)}(X_i - X_j) \tag{2.18}$$

Beim Rechnen von $\hat{h}_{bcv}$ in dieser Arbeit wird die L-BFGS-B Methode verwendet, um die BCV-Zielfunktion in folgender Form zu optimieren:

$$BCV1(k) = \frac{\prod_{l=1}^{d} k_l^{1/2}}{(2\sqrt{\pi})^d n}\left[1 + \frac{1}{32n}\sum_{i=2}^{n}\sum_{j<i}\left[\left(\sum_{l=1}^{d}(z_{ijl}k_l) - (2d+4)\right)^2 - (8d+16)\right]\prod_{l=1}^{d}(\exp(-z_{ijl}))^{k_l/4}\right] \tag{2.19}$$

und

$$BCV2(k) = \frac{\prod_{l=1}^{d} k_l^{1/2}}{(\sqrt{2\pi})^d n}\left[\frac{1}{(\sqrt{2})^d} + \frac{1}{2n}\sum_{i=2}^{n}\sum_{j<i}\left[\left(\sum_{l=1}^{d}(z_{ijl}k_l) - (d+2)\right)^2 - (2d+4)\right]\prod_{l=1}^{d}(\exp(-z_{ijl}))^{k_l/2}\right] \tag{2.20}$$

wobei $z_{ijl} = (X_{il} - X_{jl})^2$ und $k = (k_1, ...k_d)^T$ mit $k_l = 1/h_l^2$ für $l \in (1, ..., d)$.

Die Hauptkritik an der BCV Methode liegt an dem Bias des BCV Schätzers. In der Literatur wurde auch gezeigt, dass man mit der BCV Methode die unbekannte Dichtefunktion überschätzen kann. Diese Probleme werden dann in Abschnitt 2.2 anhand simulierter Daten veranschaulicht. In der praktischen Anwendung der BCV Methode besteht auch die Schwierigkeit in der Identifizierung von einem geeigneten lokalen Minimum der Zielfunktion. Darauf geht man in Abschnitt 2.3 ein.

2.1.4 DPI Direct-Plug-In

Die Direct-Plug-In Methode wurde von Sheather & Jones (1991) vorgeschlagen und von Wand & Jones (1994) auf multivariaten Fall erweitert. Das Direct-Plug-In Kriterium

basiert auf dem AMISE Kriterium und hat die gleiche Form wie bei der BCV Methode:

$$PI(h) = \frac{R(K)}{n\prod_{l=1}^{d} h_l} + \frac{1}{4}\mu_2(K)^2(h^2)^T\Psi_D(h^2) \tag{2.21}$$

Der Unterschied zwischen der DPI und BCV Methode liegt nur in der Schätzung von Ψ_D also ψ_r. Bei der DPI Methode wird ψ_r durch $\hat{\psi}_r(G)$ geschätzt, wobei G eine Pilot-Bandbreitenmatrix ist. Für eine ausführliche Beschreibung verweisen wir auf die Arbeit von Wand & Jones (1995) und Duong & Hazelton (2005). In dieser Arbeit wird die **R** Funktion Hpi.diag(ks) für die Berechnung des Direct-Plug-In Glättungsparameters $\hat{h}_{pi}$ benutzt.

Eigentlich ist der Direct-Plug-In Schätzer $\hat{h}_{pi}$ im univariaten Fall weit beliebt wegen der schnellen Konvergenzrate und kleinen Variabilität bei der Stichprobe mit endlichem Umfang. Die Hauptkritik an der DPI Methode liegt an den folgenden Punkten (Loader 1999):

1. Der DPI Schätzer hängt vom Pilot-Glättungsparameter ab. Ein großer Bias entsteht, wenn die Annahme am Pilot-Glättungsparameter nicht stimmt;

2. Bei der DPI Methode wird starke Annahme (mindestens viermal differenzierbar) an der unbekannten Zieldichte f gelegt;

3. Bei der DPI Methode wird die Kernfunktion höherer Ordnung für die Bestimmung des Pilot-Glättungsparameters verwendet.

Die Punkte 2 und 3 gelten auch für die BCV-Methode. In Abschnitt 2.2 wird die Qualität des DPI Schätzers $\hat{h}_{PI}$ anhand simulierter Daten überprüft und mit denen aus klassischen Methoden verglichen.

2.2 Vergleich der vier Methoden

In diesem Abschnitt wird im 1D, 2D und 5D Fall gezeigt, wie die vier Bandbreitenselektoren (LCV, LSCV, BCV und Direct-Plug-In) den MISE-optimalen Glättungsparameter h_{MISE} widerspiegeln. Als Beispiele verwendet man hier simulierte Daten auf Basis der **Hidalgo** Daten (Izenman & Sommer (1988)) im 1D Fall, simulierte Daten auf Basis der **geyser** Daten (Azzalini & Bowman (1990)) im 2D Fall und simulierte Daten im 5D Fall. Der hier verwendete **geyser** Datensatz ist eine Version der Eruptionsdaten aus dem „Old Faithful“ Geysir im oberen Geysir-Becken des Yellowstone-Nationalparks im Bundesstaat Wyoming (USA). Dabei gibt es 299 Messwerte von 1. August bis 15. August von zwei Merkmalen: *waiting* (Wartezeit bis zur nächsten Eruption) und *duration* (Dauer einer Eruption). Abbildung 2.1 zeigt die **geyser** Daten im Scatterplot. Die relevante Information über **Hidalgo** Daten befindet sich in Abschnitt 1.1.

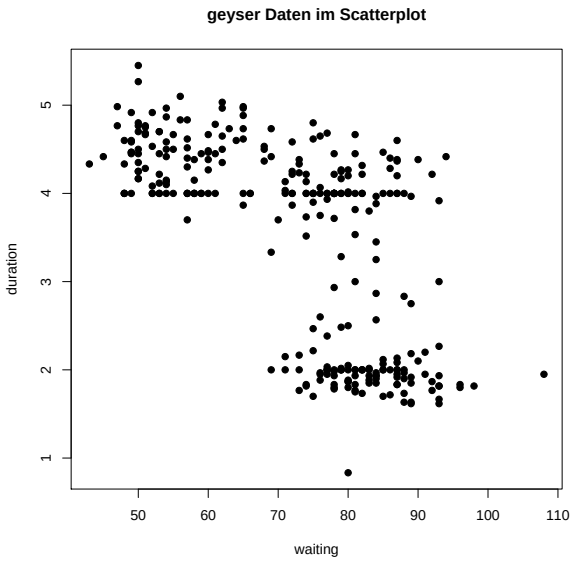

Abbildung 2.1: **geyser** Daten im Scatterplot

2.2.1 Vergleich der vier Methoden im univariaten Fall

Eine modifizierte Version der **Hidalgo** Daten (mit Hinzufügung von Störungstermen $\rho_i \sim U(-0,0005; 0,0005)$, $i = 1, ..., 485$) wird in dieser Arbeit für den Vergleich der vier Methoden im univariaten Fall verwendet. Man geht wie folgt vor:

- Man konstruiert zwei gemischte Normalverteilungen $S1$ und $S2$ aus den modifizierten Daten. Eine univariate gemischte Normaldichte definiert man wie folgt:

$$f(x) = \sum_{i=1}^{k} w_i \phi_{\sigma_i}(x - \mu_i) \tag{2.22}$$

  wobei k ein positiver Integer, $w_i > 0$ für $i = 1, ..., k$ mit $\sum_{i=1}^{k} w_i = 1$, $-\infty < \mu_i < \infty$ und $\sigma_i > 0$ für $i = 1, ..., k$. Die Parameter für die Dichtefunktionen von $S1$ und $S2$ sind hier $k = 485$, $w_1 =, ..., = w_k = \frac{1}{485}$, $\mu_i = X_i$ für $i = 1, ..., 485$, $\sigma_{i_{s1}} = (\hat{h}_{lscv} + \hat{h}_{lcv})/2 = 0,000787$ für $i = 1, ..., 485$ für die Dichtefunktion von $S1$ und $\sigma_{i_{s2}} = (\hat{h}_{bcv} + \hat{h}_{pi})/2 = 0,002837$ für $i = 1, ..., 485$ für die Dichtefunktion von $S2$, wobei $\hat{h}_{lscv}$, $\hat{h}_{lcv}$, $\hat{h}_{bcv}$ und $\hat{h}_{pi}$ die Ergebnisse aus den vier Bandbreitenselektoren auf den modifizierten Daten sind;

- Man zieht 200 Zufallsstichproben jeweils aus $S1$ und $S2$;

- Für jede Stichprobe berechnet man die Bandbreiten mit den vier Bandbreitenselektoren, d.h., man erhält dann 200 Bandbreiten für jeden Bandbreitenselektor;

- Man berechnet die MISE-optimale Bandbreite h_{MISE}. Wenn man den Produkt-Normalkern bei Kerndichteschätzung verwendet, dann hat die Funktion $MISE\{\hat{f}(x;h)\}$ eine geschlossene Form wie folgt:

$$MISE\left\{\hat{f}(x;h)\right\} = \frac{1}{2\pi^{1/2}nh} + w^T((1-\frac{1}{n})\Omega_2 - 2\Omega_1 + \Omega_0)w \tag{2.23}$$

  wobei $w = (w_1, ..., w_k)^T$, und Ω_a eine $k \times k$ Matrix mit dem (i,j)-ten Element $\Omega_a(i,j) = \phi_{(ah^2+\sigma_i^2+\sigma_j^2)}(\mu_i - \mu_j)$ für $i,j \in (1,...,k)$. Man kann numerische Methode verwenden, um das Minimum der Funktion $MISE\{\hat{f}(x;h)\}$, also die MISE-optimale Bandbreite zu finden. Mit **R** Funktion optimize(stats) bekommt man die zwei MISE-optimalen Bandbreiten $h_{MISE1} = 0,000791$ für den Kerndichteschätzer der Dichtefunktion von $S1$ und $h_{MISE2} = 0,001957$ für den Kerndichteschätzer der Dichtefunktion von $S2$;

- Man fasst die Resultate zusammen und stellt sie grafisch dar. Bei der grafischen Darstellung wird h_{MISE} (hier h_{MISE1} und h_{MISE2}) in senkrechter Linie in lila markiert. Um die Verteilung der 200 Bandbreiten für jeden Bandbreitenselektor besser zu erkennen, schätzt man die Dichtefunktion der Bandbreiten mit **R** Funktion density(stats) mit der Normal-Reference Bandbreite und zeichnet die Kurve dieses Dichteschätzers mit in der Grafik.

In der Grafik oben links in Abbildung 2.2 und 2.3 werden die Dichtefunktionen von $S1$ und $S2$ im Contour Plot dargestellt. Abbildung 2.2 zeigt die Resultate aus den 200 Stichproben aus $S1$, wobei die Grafik oben rechts/mittel links/mittel rechts/unten links/unten rechts für $\hat{h}_{lcv}/\hat{h}_{lscv}/\hat{h}_{bcv1}/\hat{h}_{bcv2}/\hat{h}_{pi}$ steht. Analog zeigt Abbildung 2.3 die Resultate aus den 200 Stichproben aus $S2$. Wenn man die Resultate in Abbildung 2.2 und 2.3 vergleicht, dann ist der folgende Schluss zu ziehen:

- LSCV bringt die beste Leistung: $\hat{h}_{lscv}$ hat zwar größere Variabilität als $\hat{h}_{pi}$, aber dafür einen kleinen Bias zu h_{MISE};

- BCV2 bringt die zweite beste Leistung: $\hat{h}_{bcv1}$ hat einen kleinen Bias zu h_{MISE}, aber größere Variabilität als $\hat{h}_{pi}$ und $\hat{h}_{lscv}$;

- $\hat{h}_{pi}$ hat kleine Variabilität aber einen großen Bias zu h_{MISE}. Die DPI Methode liefert eigentlich ein schlechtes Resultat trotz der kleinen Variabilität;

- $\hat{h}_{lcv}$ und $\hat{h}_{bcv1}$ haben sowohl große Variabilität als auch einen großen Bias zu h_{MISE}, d.h., die zwei Methoden liefern die schlechtesten Ergebnisse.

Zu bemerken ist, dass man hier beim Anwenden der BCV Methode auf den 200 Stichproben aus $S1$ im Unterschied zu der Empfehlung der Literatur (Details vgl. Abs. 2.3) das kleinere lokale Minimum genommen hat, falls es mehrere lokale Minima im gesuchten Bereich gibt. Ansonsten liefern $BCV1$ und $BCV2$ in diesem Fall relativ schlechtere Resultate, wie man in Abbildung 2.4 sieht.

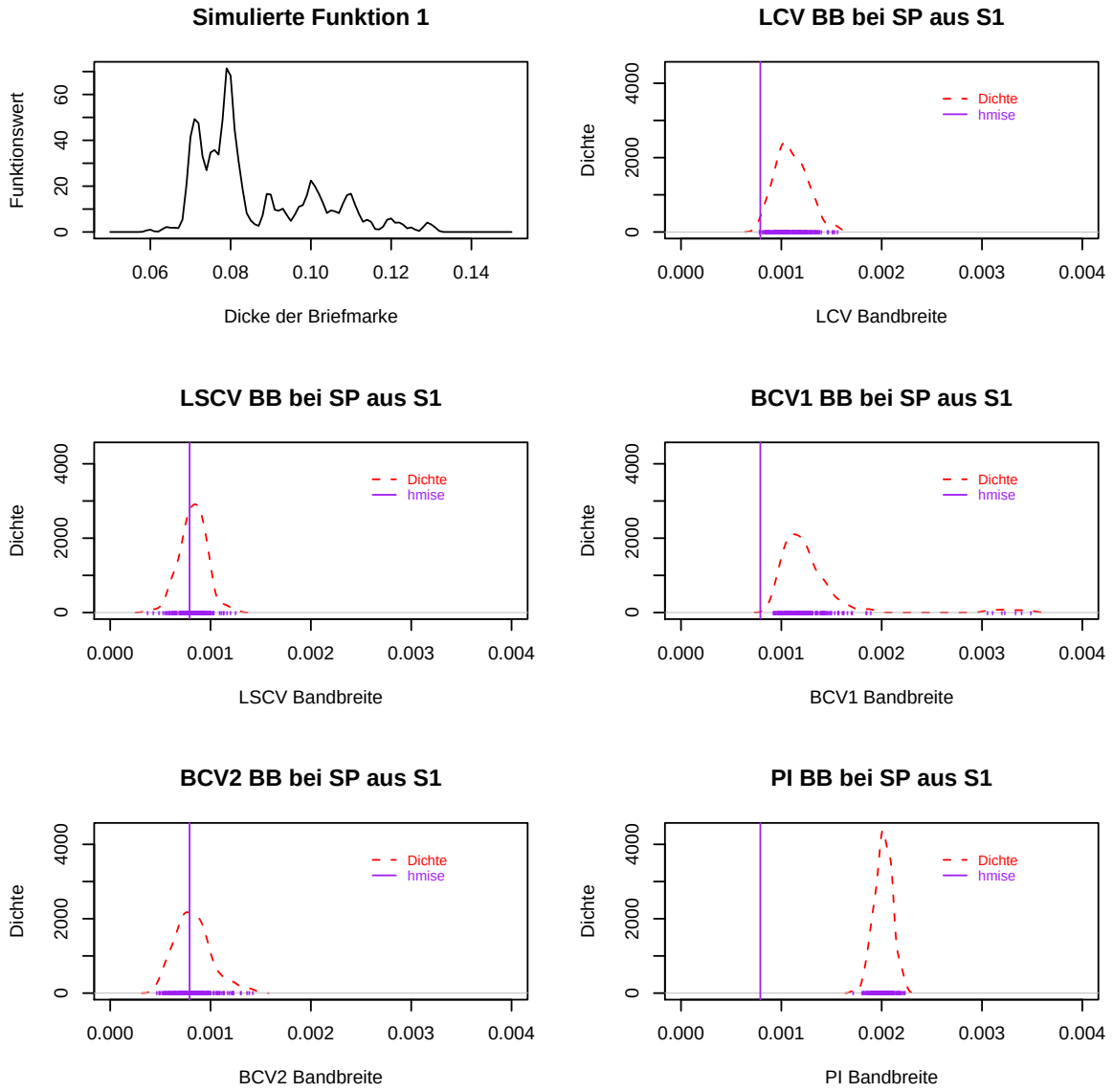

Abbildung 2.2: Dichtefunktion von $S1$ in der Grafik oben links. Verteilung der Bandbreiten aus vier Bandbreitenselektoren (LCV, LSCV, BCV (BCV1 und BCV2) und DPI) bei 200 Stichproben aus $S1$: Grafik oben rechts/mittel links/mittel rechts/unten links/unten rechts steht für $\hat{h}_{lcv}/\hat{h}_{lscv}/\hat{h}_{bcv1}/\hat{h}_{bcv2}/\hat{h}_{pi}$.

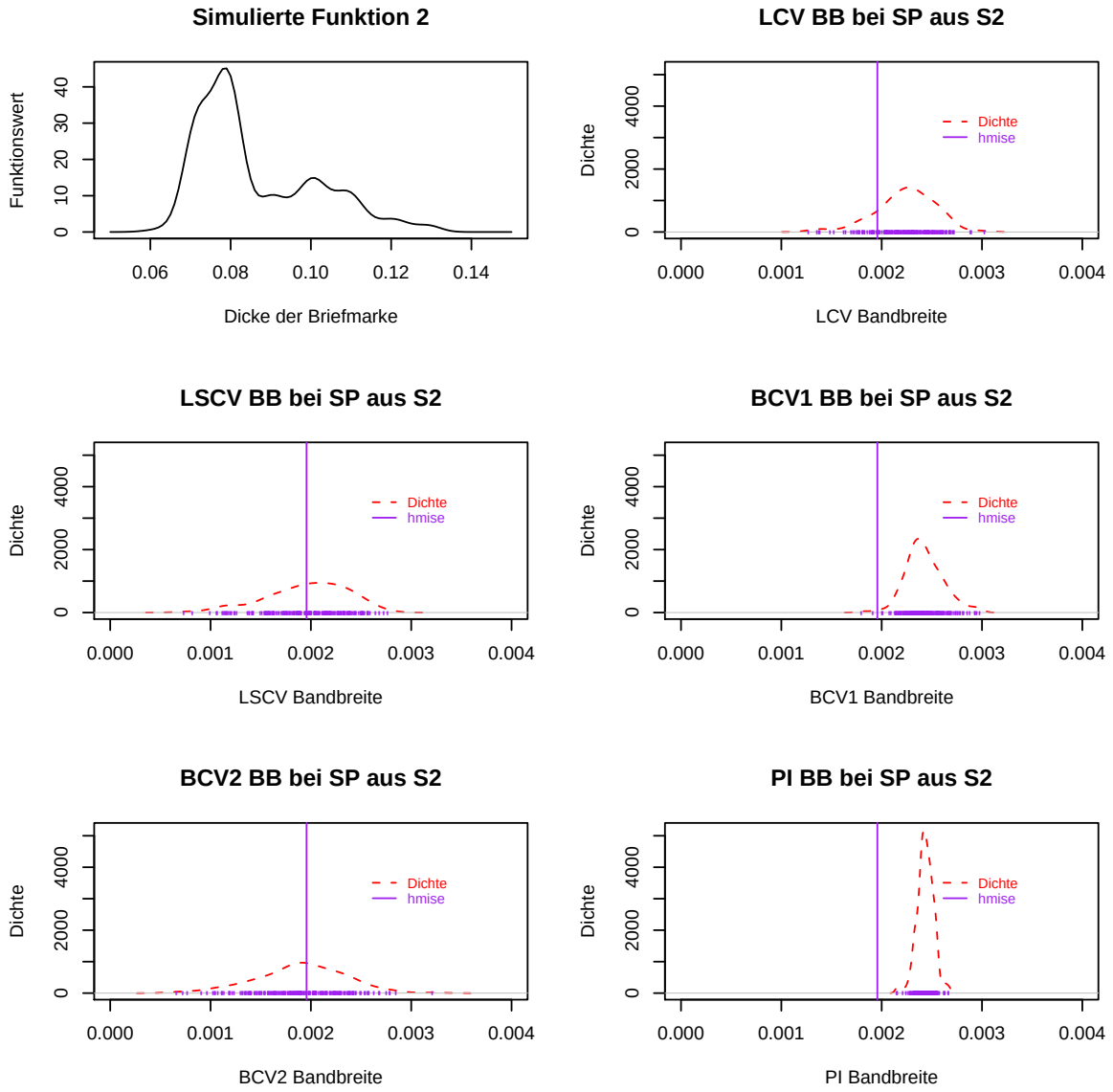

Abbildung 2.3: Dichtefunktion von $S2$ in der Grafik oben links. Verteilung der Bandbreiten aus vier Bandbreitenselektoren (LCV, LSCV, BCV (BCV1 und BCV2) und DPI) bei 200 Stichproben aus $S2$: Grafik oben rechts/mittel links/mittel rechts/unten links/unten rechts steht für $\hat{h}_{lcv}/\hat{h}_{lscv}/\hat{h}_{bcv1}/\hat{h}_{bcv2}/\hat{h}_{pi}$.

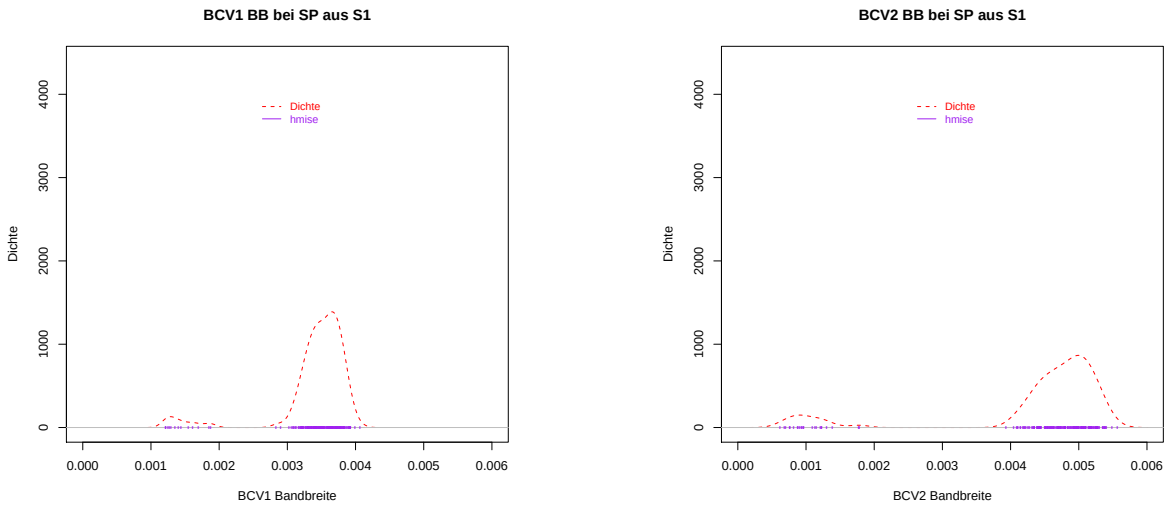

Abbildung 2.4: Verteilung der Bandbreiten aus BCV1 (Grafik links) und BCV2 (Grafik rechts) bei 200 Stichproben aus $S1$, wenn man das größte lokale Minimum nimmt.

2.2.2 Vergleich der vier Methoden im bivariaten Fall

Eine modifizierte Version der **geyser** Daten (mit Hinzufügung von $\rho_i \sim U(-0,008333; 0,008333)$, $i = 1, ..., 299$ zur Variable *duration*) wird in dieser Arbeit für den Vergleich der vier Methoden im bivariaten Fall verwendet. Beim Vergleich der vier Methoden im multivariaten Fall geht man genauso wie im univariaten Fall vor.

Eine multivariate gemischte Normaldichte hat die folgende Form:

$$f(x) = \sum_{i=1}^{k} w_i \phi_{\Sigma_i}(x - \mu_i) \tag{2.24}$$

wobei k ein positiver Integer, $w = (w_1, ..., w_k)^T$ ein Vektor mit positiven Elementen und $\sum_{i=1}^{k} w_i = 1$, μ_i ein $d \times 1$ Vektor und Σ_i die $d \times d$ Kovarianzmatrix für $i = 1, ..., k$ (Wand & Jones (1995)). In diesem Kapitel ist Σ_i eine Diagonalmatrix mit diagonalen Elementen $\sigma_{i1}^2, ..., \sigma_{id}^2$.

Man konstruiert hier zwei gemischte Normaldichten wie folgt:

$$f_{2d1} = \sum_{i=1}^{k} w_i \phi_{\Sigma_{i1}}(x - \mu_i) \tag{2.25}$$

und

$$f_{2d2} = \sum_{i=1}^{k} w_i \phi_{\Sigma_{i2}}(x - \mu_i) \tag{2.26}$$

wobei $k = 299$, $w_1 =, ..., = w_k = \frac{1}{299}$, und $\mu_i = (X_{i1}, X_{i2})^T$ für $i = 1, ..., k$ aus den modifizierten **geyser** Daten, $\Sigma_{i1} = \begin{pmatrix} 1,0076 & 0 \\ 0 & 0,0205 \end{pmatrix}$ und $\Sigma_{i2} = \begin{pmatrix} 3,3149 & 0 \\ 0 & 0,2375 \end{pmatrix}$ für $i = 1, ..., k$. In der Grafik oben links in Abbildung 2.5 und 2.6 werden f_{2d1} und f_{2d2} im Contour Plot dargestellt. Wenn man im multivariaten Fall den Produkt-Normalkern bei Kerndichteschätzung verwendet, dann hat $MISE\{\hat{f}(x;h)\}$ mit $h = (h_1, ..., h_d)^T$ unter gemischter Normaldichte f eine geschlossene Form wie folgt

$$MISE\left\{\hat{f}(x;h)\right\} = \frac{1}{n(4\pi)^{d/2}\prod_{l=1}^{d} h_l} + w^T \left\{(1 - \frac{1}{n})\Omega_2 - 2\Omega_1 + \Omega_0\right\} w \tag{2.27}$$

wobei Ω_a eine $k \times k$ Matrix, deren (i,j)-tes Element

$$\Omega_a(i,j) = \prod_{l=1}^{d} \phi_{ah_l^2 + \sigma_{il}^2 + \sigma_{jl}^2}(\mu_{il} - \mu_{jl}) \tag{2.28}$$

für $i, j \in \{1, ..., k\}$, falls Σ_i, $i \in \{1, ..., k\}$ eine Diagonalmatrix ist. Aufgrund von (2.27) und (2.28) kann man numerische Methode verwenden, um das lokale Minimum von $MISE\{\hat{f}(x;h)\}$ zu bestimmen, das der MISE-optimalen Bandbreite h_{MISE} entspricht. Mit der quasi-Newton Methode bekommt man $h_{MISE1} = \begin{pmatrix} 2,1742 & 0 \\ 0 & 0,1294 \end{pmatrix}$ für den Kerndichteschätzer für f_{2d1} und $h_{MISE2} = \begin{pmatrix} 2,9495 & 0 \\ 0 & 0,2980 \end{pmatrix}$ für den Kerndichteschätzer für f_{2d2}.

Analog wie in 2.2.1 zeigt Abbildung 2.5 die Resultate aus den 200 Stichproben mit Dichtefunktion f_{2d1}, wobei die Grafik oben rechts/mittel links/mittel rechts/unten links/unten rechts für $\hat{h}_{lcv}/\hat{h}_{lscv}/\hat{h}_{bcv1}/\hat{h}_{bcv2}/\hat{h}_{pi}$ steht. Als Referenz wird die MISE-optimale Bandbreite h_{MISE1} mit einem grünen Kreuz markiert. Um die Verteilung der Bandbreiten besser zu erkennen, wird die Dichte der Bandbreiten mit **R** Funktion kde(ks) mit der Direct-Plug-In Bandbreite geschätzt und in den Grafiken in roten Contour Linien gezeichnet. Analog zeigt Abbildung 2.6 die Resultate aus den 200 Stichproben mit Dichtefunktion f_{2d2}. Aus dem Vergleich der Resultate in Abbildung 2.5 und 2.6 ist der folgende Schluss zu ziehen:

- Die LSCV-Bandbreite hat einen kleinen Bias zu h_{MISE} in beiden Fällen;
- $\hat{h}_{pi}$ hat in beiden Fällen die kleinste Variabilität;
- Die LCV Methode bringt eine ziemlich gut Leistung bei Stichproben mit Dichte-

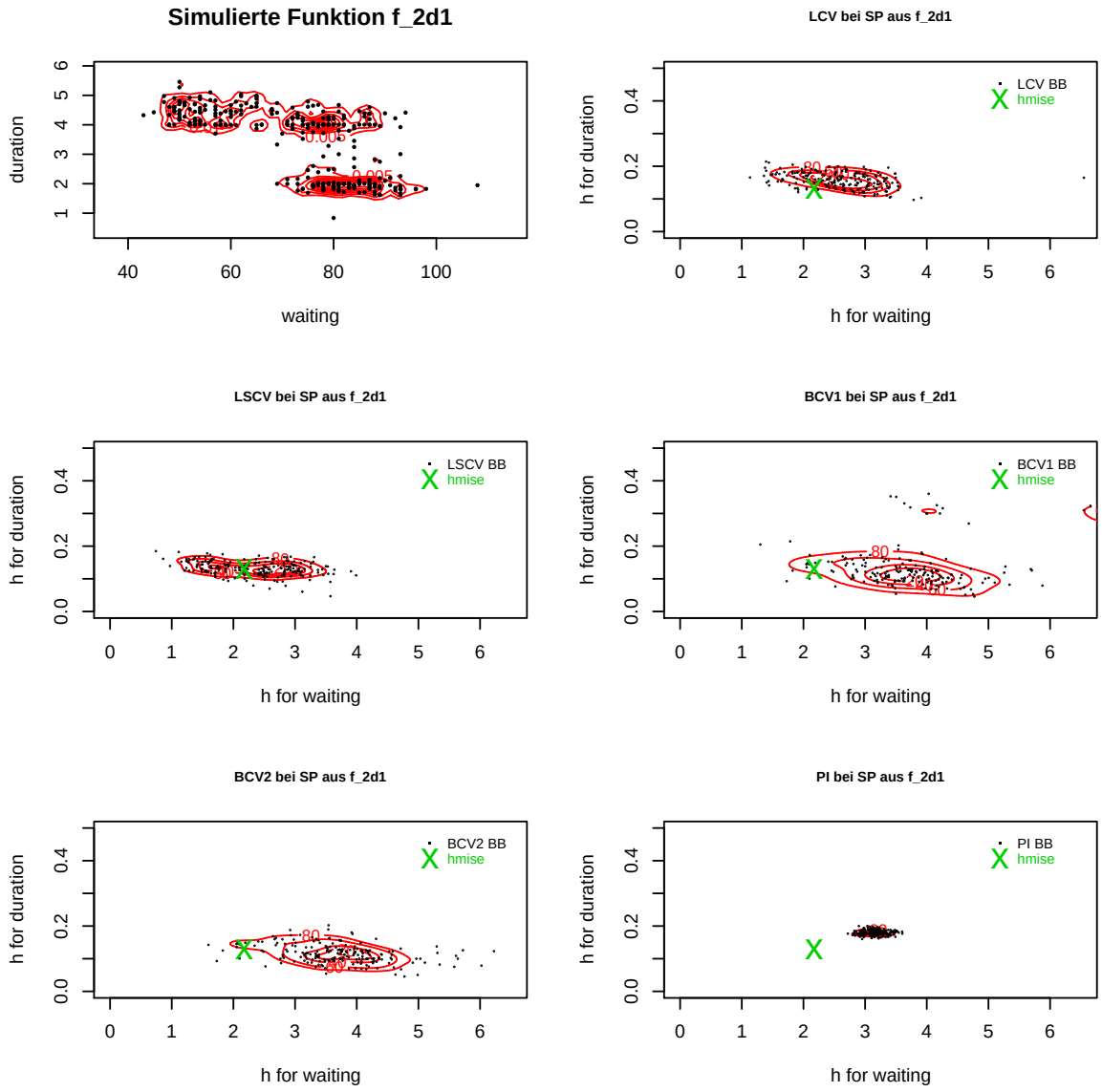

Abbildung 2.5: Dichtefunktion f_{2d1} im Contour Plot in der Grafik oben links. Verteilung der Bandbreiten aus vier Bandbreitenselektoren (LCV, LSCV, BCV (BCV1 und BCV2) und DPI) bei 200 Stichproben mit Dichtefunktion f_{2d1}: Grafik oben rechts/mittel links/mittel rechts/unten links/unten rechts steht für $\hat{h}_{lcv}/\hat{h}_{lscv}/\hat{h}_{bcv1}/\hat{h}_{bcv2}/\hat{h}_{pi}$.

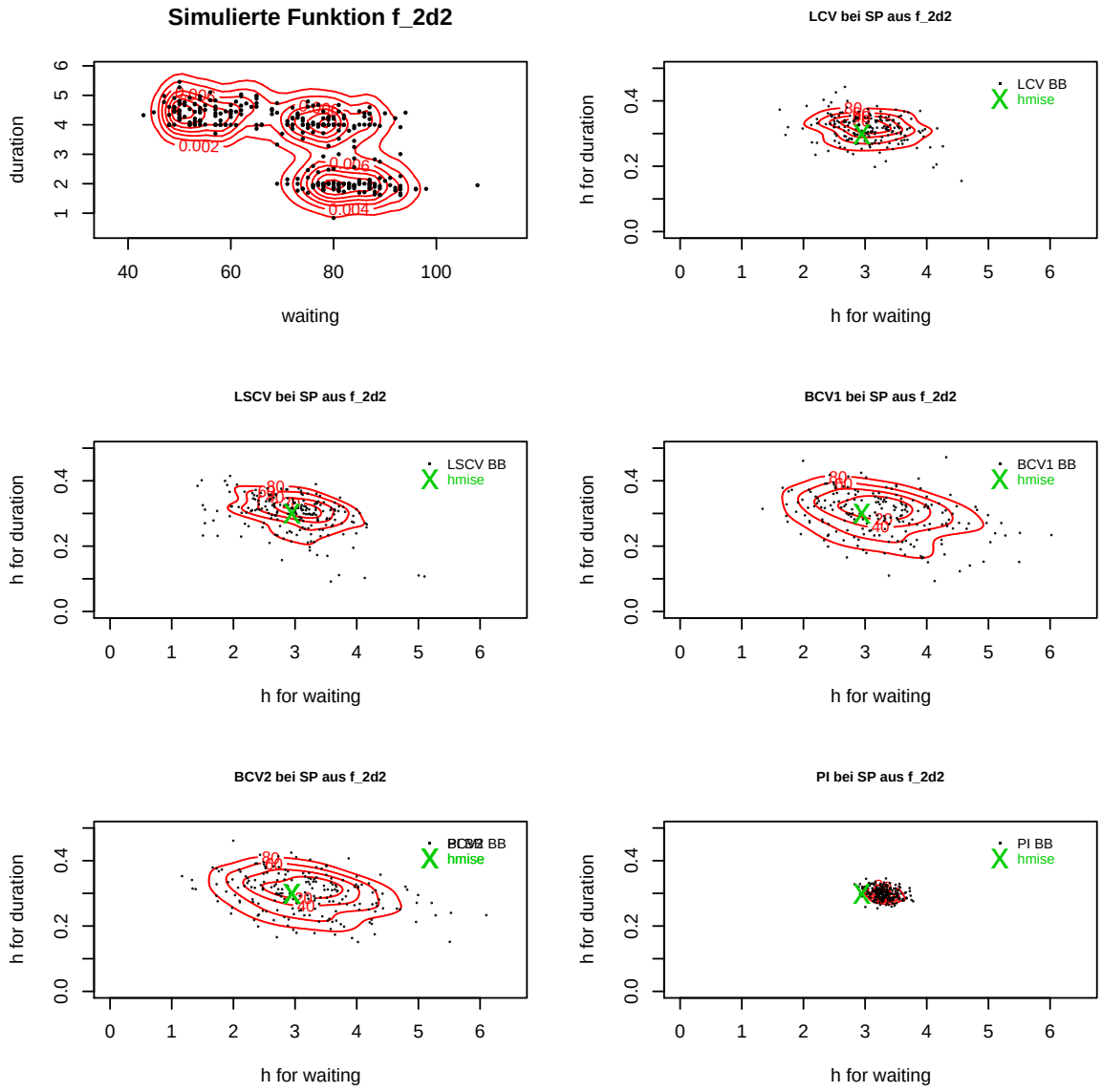

Abbildung 2.6: Dichtefunktion f_{2d2} im Contour Plot in der Grafik oben links. Verteilung der Bandbreiten aus vier Bandbreitenselektoren (LCV, LSCV, BCV (BCV1 und BCV2) und DPI) bei 200 Stichproben mit Dichtefunktion f_{2d2}: Grafik oben rechts/mittel links/mittel rechts/unten links/unten rechts steht für $\hat{h}_{lcv}/\hat{h}_{lscv}/\hat{h}_{bcv1}/\hat{h}_{bcv2}/\hat{h}_{pi}$.

funktion f_{2d2};

- Der Bias von $\hat{h}_{bcv}$ und $\hat{h}_{pi}$ zu h_{MISE} hängt stark von der Varianz der Daten ab. Sie neigen dazu, die unbekannte Dichtefunktion zu überschätzen;
- Die BCV Methoden (BCV1 und BCV2) liefern in beiden Fällen eine schlechte Schätzung von h_{MISE} wegen ihres großen Bias zu h_{MISE} und der großen Variabilität.

2.2.3 Vergleich der vier Methoden im 5D Fall

Für den Vergleich der vier Bandbreitenselektoren im 5D Fall werden 200 Stichproben mit jeweils 300 Daten aus der folgenden gemischten Normaldichte simuliert:

$$f_{5d}(x) = \frac{1}{3}\sum_{i=1}^{3}\phi_D(x-\mu_i) \tag{2.29}$$

wobei D eine 5D Einheitsmatrix, $\mu_1 = (1,1,1,1,0)^T$, $\mu_2 = (1,1,1,0,1)^T$ und $\mu_3 = (1,1,0,1,1)^T$. Analog wie in 2.2.2 berechnet man $h_{MISE} = (0,5517; 0,5514; 0,6107; 0,6102; 0,6101)^T$. Die Resultate aus den vier Bandbreitenselektoren werden in Abbildung 2.7 im Parallel Koordinaten Plot dargestellt. Um die Verteilung der Bandbreiten besser zu erkennen, wird die Dichte der Bandbreiten mit **R** Funktion kde(ks) mit der Direct-Plug-In Bandbreite geschätzt und in die Farbe der Linien abgebildet, wobei die Farbpalette gray.colors(10) verwendet wird, mit der die dunklen Linien im Parallel Koordinaten Plot für die Bandbreiten aus High Density Regions stehen. Als Referenz markiert man h_{MISE} mit einer grünen Linie. Abbildung 2.8-2.12 zeigen die Scatterplot Matrizen von den 2D Projektionen der 5D Bandbreiten aus den vier Bandbreitenselektoren, wobei die 2D Projektionen des 5D-Kerndichteschätzers mit der Direct-Plug-In Bandbreite in roten Contour Linien mit gezeichnet werden, um die Verteilung der Bandbreiten besser zu erkennen. Als Referenz werden die 2D Projektionen von h_{MISE} mit grünem Kreuz markiert. Kurz zu erwähnen ist, dass die Projektionen in der 4×5D Ebene der Gleichmäßigkeit der grafischen Darstellung halber nicht gezeigt werden. In Abbildung 2.7-2.12 sieht man folgendes:

- Die BCV1 Methode zeigt eine bessere Leistung als die im uni- und bivariaten Fall: $\hat{h}_{bcv1}$ hat kleine Variabilität und auch keinen großen Bias;
- Die LCV Methode bringt eine gute Leistung;
- $\hat{h}_{lscv}$ hat zwar einen kleinen Bias, aber dafür große Variabilität;
- Die DPI Methode unterschätzt die MISE-optimale Bandbreite im 5D Fall. Der Grund liegt vermutlich an der Verwendung der SAMSE (Sum of Asymptotic Mean Squared Error) Pilot-Bandbreite. $\hat{h}_{pi}$ hat zwar kleine Variabilität, aber dafür einen großen Bias;

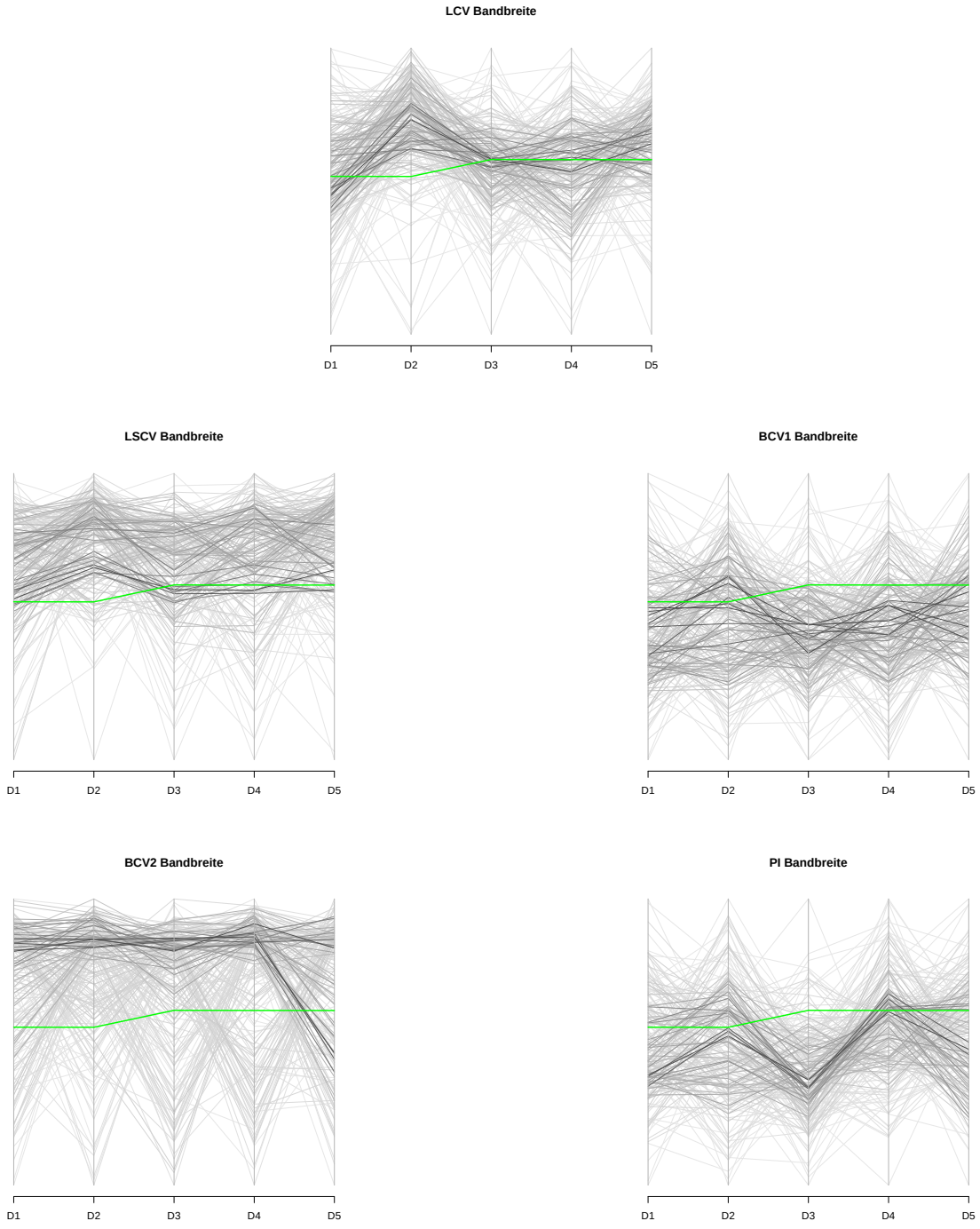

Abbildung 2.7: Verteilung der Bandbreiten aus vier Bandbreitenselektoren (LCV, LSCV, BCV (BCV1 und BCV2) und DPI) bei den 200 Stichproben mit Dichtefunktion f_{5d} im 5D Fall im Parallel Koordinaten Plot: Grafik oben/mittel links/mittel rechts/unten links/unten rechts steht für $\hat{h}_{lcv}/\hat{h}_{lscv}/\hat{h}_{bcv1}/\hat{h}_{bcv2}/\hat{h}_{pi}$.

- Die BCV2 Methode ist in diesem Beispiel wegen der großen Verzerrung und Variabilität nicht mehr zuverlässig.

In diesem Abschnitt werden vier Bandbreitenselektoren (LCV, LSCV, BCV und Direct-Plug-In) anhand simulierter Daten verglichen. Es hat sich ergeben, dass keine Methode in der Lage ist, einen MISE-optimalen Glättungsparameter zu liefern, was auch von dem Umfang der Stichprobe abhängt. In diesem Sinne könnte man in der explorativen Datenanalyse nützliche Information über die Daten verlieren, wenn man nur mit einem theoretisch gesehen optimalen Glättungsparameter arbeitet. Aus diesem Grund ist in der Praxis stark zu empfehlen, die unbekannte Dichte der Daten anhand verschiedener Glättungsparameter zu schätzen, damit man die Daten aus unterschiedlichen Aspekten betrachten kann.

2.3 Diskussion über die Probleme bei CV Methoden

2.3.1 LSCV bei Daten mit diskretisierten Variablen

Wie oben erwähnt wird die Zielfunktion $LSCV(h)$ wie folgt geschrieben:

$$LSCV(h) = \int_{R^d} \hat{f}(x;h)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i;h) \tag{2.30}$$

und falls ein Normalkern bei Kerndichteschätzung verwendet wird, dann gilt:

$$LSCV(h) = \frac{1}{(n-1)}\prod_{l=1}^{d}\phi(0;\sqrt{2}h_l) + \frac{n-2}{n(n-1)^2}\sum_{i\neq j}\prod_{l=1}^{d}\phi(X_{il}-X_{jl};\sqrt{2}h_l) - \frac{2}{n(n-1)}\sum_{i\neq j}\prod_{l=1}^{d}\phi(X_{il}-X_{jl};h_l) \tag{2.31}$$

mit $i, j \in \{1, ..., n\}$. Der Einfachheit halber kann man $n \pm 1$ in der obigen Formel durch n ersetzen (Scott 1992). Wenn man $(X_{il} - X_{jl})^2 = (X_{jl} - X_{il})^2$ berücksichtigt, dann bekommt man mit Umformung von (2.31) die folgende vereinfachte Form von $LSCV(k)$, $k = (k_1, ...k_d)^T$ mit $k_l = 1/h_l^2$ für $l \in \{1, ..., d\}$:

$$LSCV(k) = \frac{\prod_{l=1}^{d} k_l^{1/2}}{n(\sqrt{2\pi})^d}\left(\frac{1}{2^{d/2}} + \frac{n-1}{2^{d+3}} + \frac{1}{n}\sum_{j<i}^{i=2,...,n}\left[-4\left(\prod_{l=1}^{d} Z_{ijl}^{k_l} - \frac{1}{4\cdot 2^{d/2}}\right)^2\right]\right) \tag{2.32}$$

wobei $Z_{ijl} = \exp\{-(X_{il} - X_{jl})^2/4\}$.

Das Verhalten von $LSCV(k)$ ist schwer festzustellen, weil das von Z_{ijl} abhängt. Insbesondere hat Z_{ijl} einen großen Einfluss auf $LSCV(k)$, wenn $\exists l$, $k_l \to \infty$, weil in diesem Fall $Z_{ijl}^{k_l} \to 0$ für $Z_{ijl} \neq 1$ und $Z_{ijl}^{k_l} = 1$, falls $Z_{ijl} = 1$, also $X_{il} = X_{jl}$. Im Folgenden wird anhand praktischer Daten gezeigt, wie diskretisierte Variablen in den Daten die Funktion $LSCV(k)$ beeinflussen. Es wird sich folgendes ergeben:

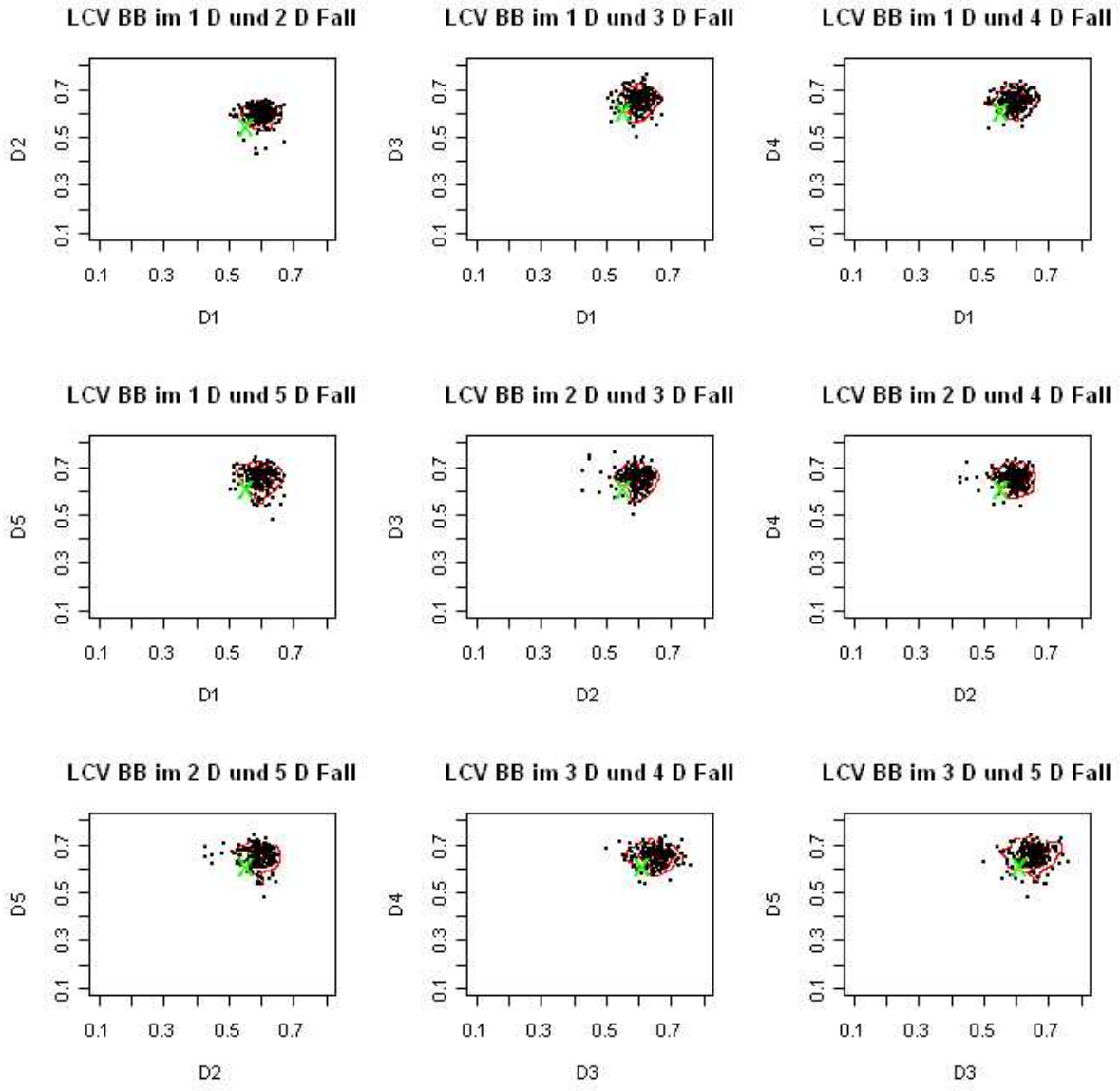

Abbildung 2.8: 2D Projektionen der 5D LCV-Bandbreiten aus 200 Stichproben mit Dichtefunktion f_{5d}. 2D Projektionen des 5D-Kerndichteschätzers mit der Direct-Plug-In Bandbreite in roten Contour Linien. 2D Projektionen von h_{MISE} in grünem Kreuz.

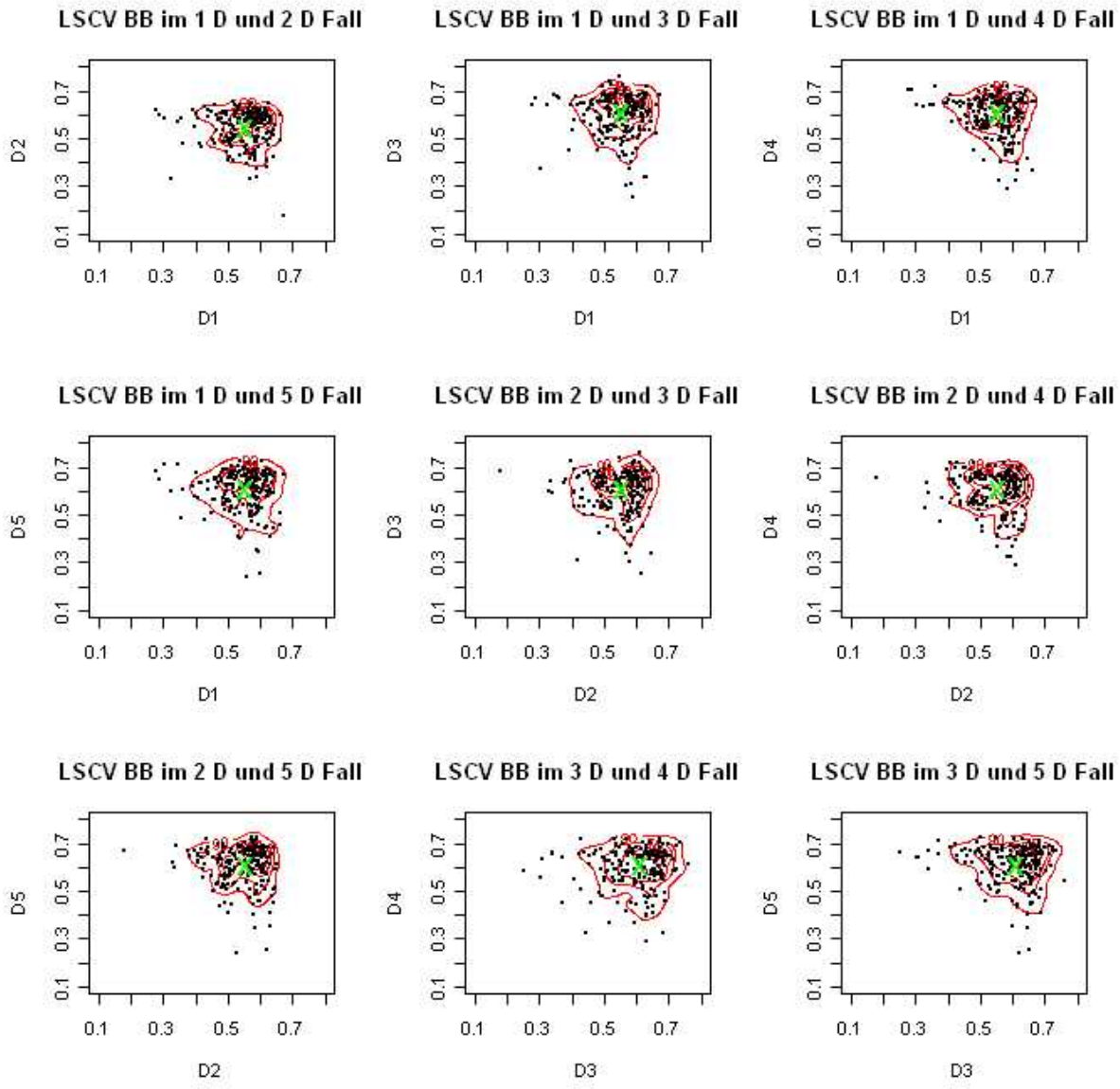

Abbildung 2.9: 2D Projektionen der 5D LSCV-Bandbreiten aus 200 Stichproben mit Dichtefunktion f_{5d}. 2D Projektionen des 5D-Kerndichteschätzers mit der Direct-Plug-In Bandbreite in roten Contour Linien. 2D Projektionen von h_{MISE} in grünem Kreuz.

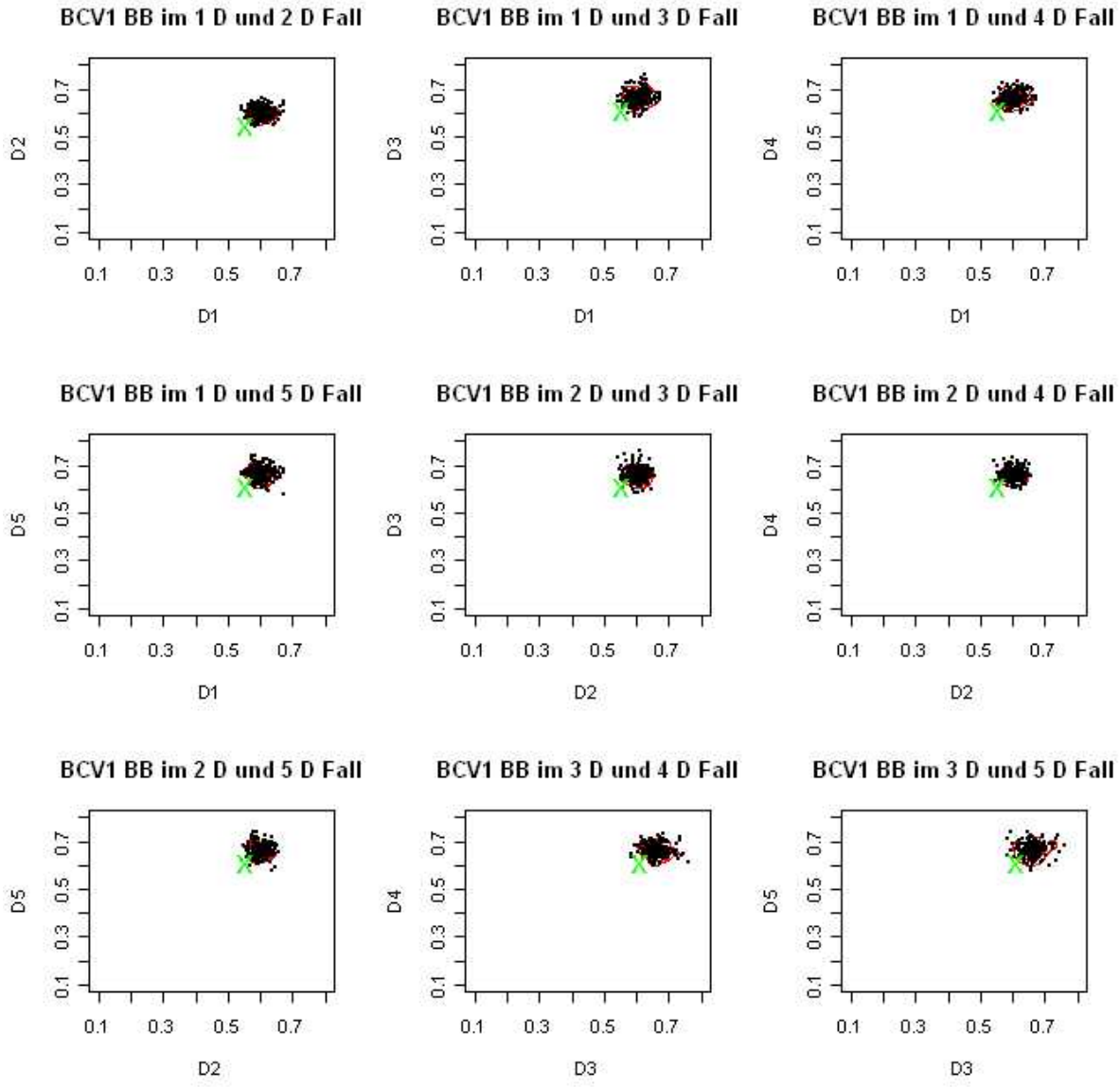

Abbildung 2.10: 2D Projektionen der 5D BCV1-Bandbreiten aus 200 Stichproben mit Dichtefunktion f_{5d}. 2D Projektionen des 5D-Kerndichteschätzers mit der Direct-Plug-In Bandbreite in roten Contour Linien. 2D Projektionen von h_{MISE} in grünem Kreuz.

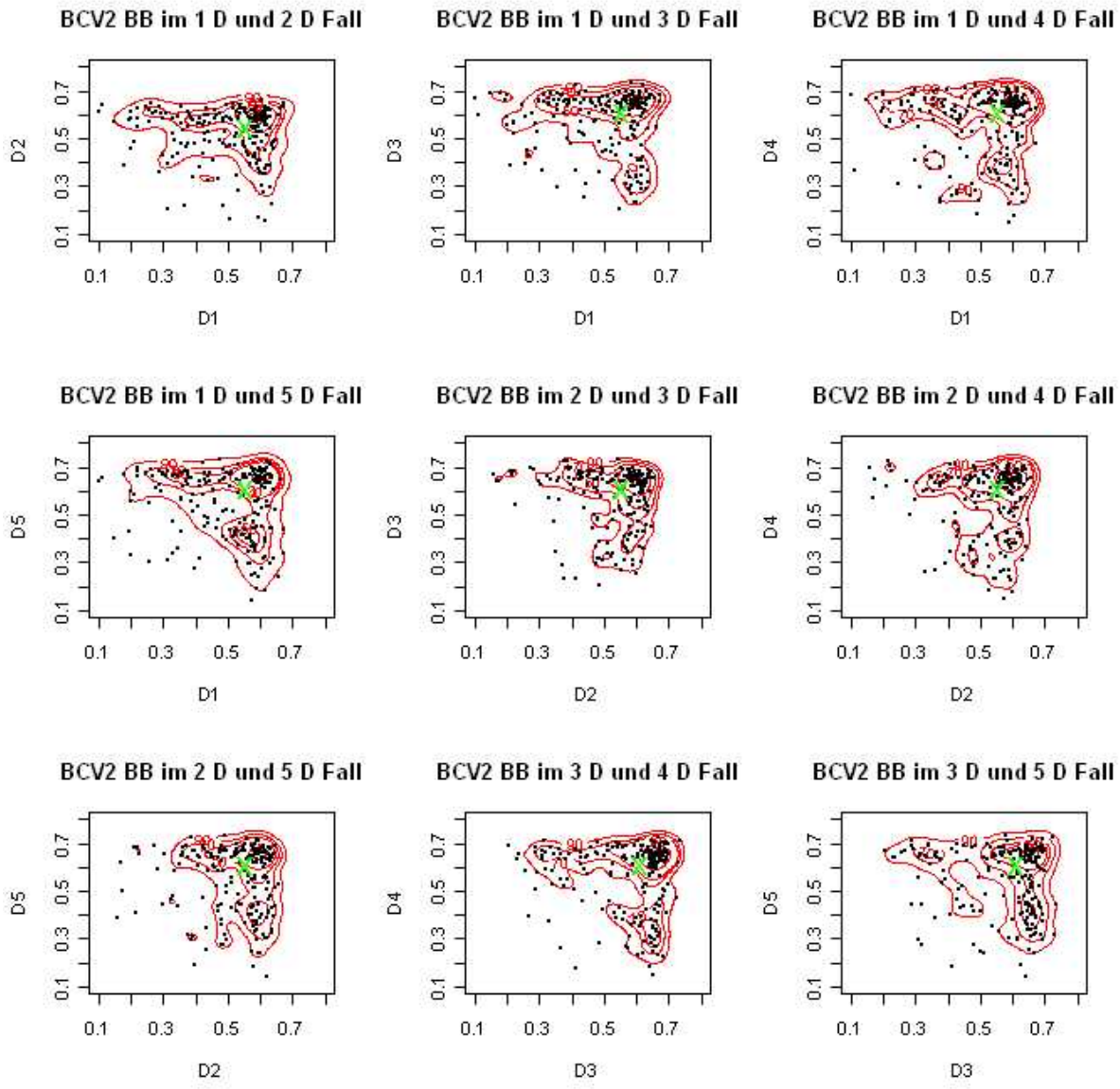

Abbildung 2.11: 2D Projektionen der 5D BCV2-Bandbreiten aus 200 Stichproben mit Dichtefunktion f_{5d}. 2D Projektionen des 5D-Kerndichteschätzers mit der Direct-Plug-In Bandbreite in roten Contour Linien. 2D Projektionen von h_{MISE} in grünem Kreuz.

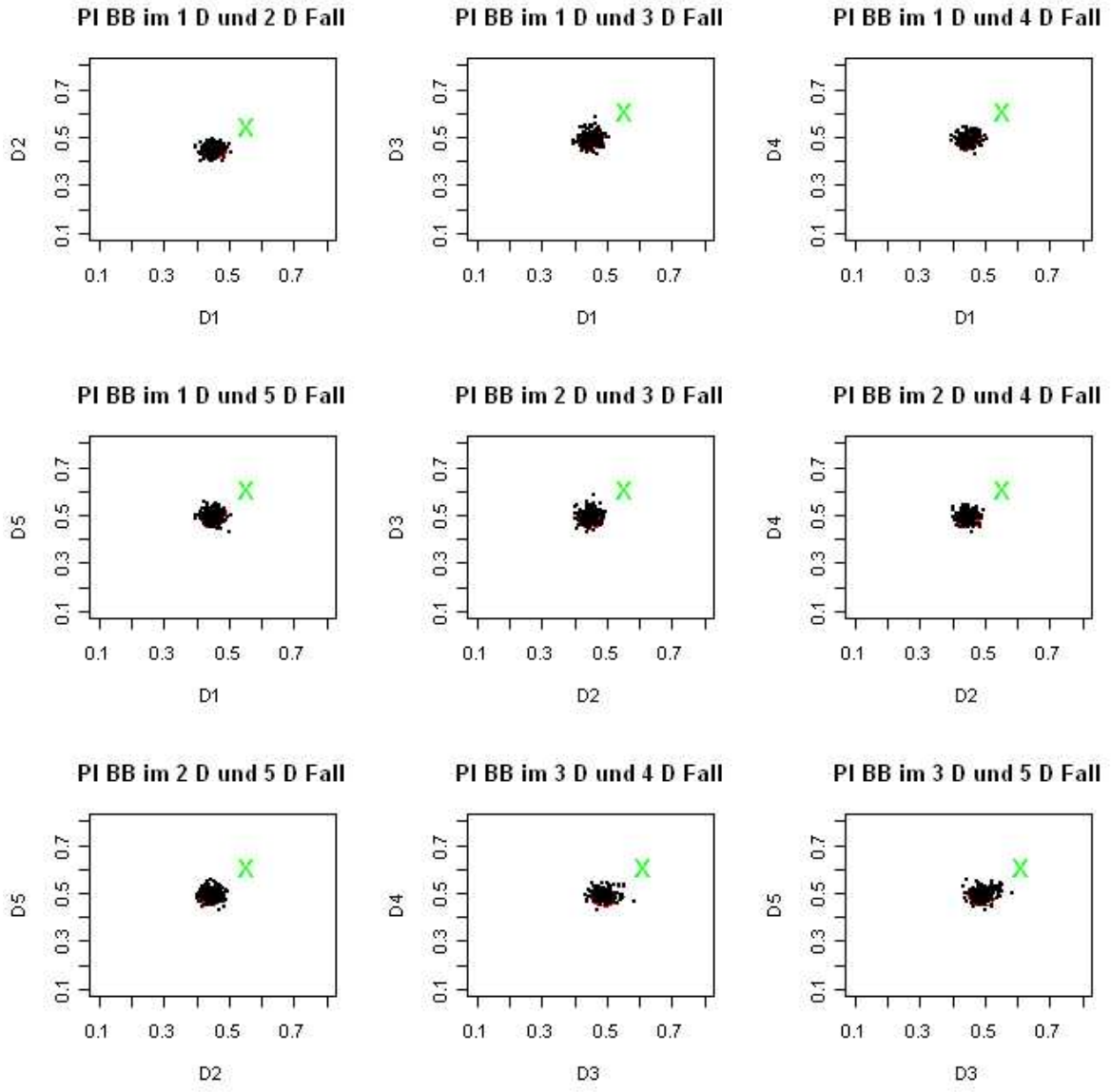

Abbildung 2.12: 2D Projektionen der 5D DPI-Bandbreiten aus 200 Stichproben mit Dichtefunktion f_{5d}. 2D Projektionen des 5D-Kerndichteschätzers mit der Direct-Plug-In Bandbreite in roten Contour Linien. 2D Projektionen von h_{MISE} in grünem Kreuz.

- Bei Daten mit diskretisierten Variablen kann $LSCV(k)$ gegen minus unendlich gehen, wenn $\exists l$, $k_l \to \infty$, aber das bedeutet **nicht**, dass man kein lokales Minimum in dem gesuchten Intervall finden kann, was in der Literatur oft ignoriert worden ist;
- Das Verhalten von $LSCV(k)$ ist bei Daten mit mehreren diskretisierten Variablen ziemlich kompliziert.

2.3.1.1 $LSCV(k)$ im univariaten Fall

Sei m die Anzahl von Bindungen $X_i = X_j$, $j < i$ in den Daten, dann

$$LSCV(k) = \frac{k^{1/2}}{n\sqrt{2\pi}} \left(\frac{1}{2^{1/2}} + \frac{m(8\sqrt{2} - 32)}{8n} \right) \tag{2.33}$$

wenn $k \to \infty$. $LSCV(k)$ wird gegen minus unendlich gehen, wenn der Term in Klammern kleiner Null ist. Durch Lösung der Ungleichung bekommt man $m > n/(4\sqrt{2} - 2)$, was mit dem Ergebnis mit einer trivialen Kernfunktion von Silverman (1986) $m > n \cdot K * K(0)/(4K(0) - 2K * K(0))$ übereinstimmt. In der Literatur wurde dann manchmal behauptet, dass $LSCV(h)$ in diesem Fall eine triviale Lösung $\hat{h}_{lscv} = 0$ nimmt. Generell gilt die Behauptung nicht, weil man in manchen Fällen ein anderes lokales Minimum als $\hat{h}_{lscv}$ nehmen kann, weil das größte Minimum beim Anwenden der LSCV-Methode ausgewählt werden soll (Wand & Jones (1995)). Das folgende kleine Beispiel zeigt, dass man trotz des schlechten Verhaltens von $LSCV(h)$ für h sehr klein eine LSCV-optimale Bandbreite erhalten kann. Zum besseren Verständnis vom Glättungsparameter wird $LSCV(h)$ anstatt von $LSCV(k)$ in dem Beispiel benutzt.

Beispiel 2.3.1: geyser Datensatz

Hier verwendet man den Kerndichteschätzer mit Glättungsparameter $\hat{h}_{lscv}$, um die unbekannte Dichte von *waiting* zu schätzen. Das Verhalten von Zielfunktion $LSCV(h)$ wird sehr schlecht, wenn h gegen Null geht, weil $m = 1095 >> n/(4\sqrt{2} - 2) = 81,7643$. Trotzdem findet man das lokale Minimum an der Stelle etwa 2,2047. Abbildung 2.13 zeigt den Verlauf von $LSCV(h)$ auf $h = [1, 3]$ (Grafik links) und den Kerndichteschätzer mit Glättungsparameter $\hat{h}_{lscv} = 2,2047$ (Grafik rechts).

2.3.1.2 $LSCV(k)$ im bivariaten Fall

Seien M_1, M_2 und M_{12} die Indexmengen von Bindungen mit

$$M_1 = ((i,j)|X_{i1} = X_{j1},\ j < i)\,,$$

$$M_2 = ((i,j)|X_{i2} = X_{j2},\ j < i)$$

und

$$M_{12} = ((i,j)|X_{i2} = X_{j2} \wedge X_{i2} = X_{j2},\ j < i)\,,$$

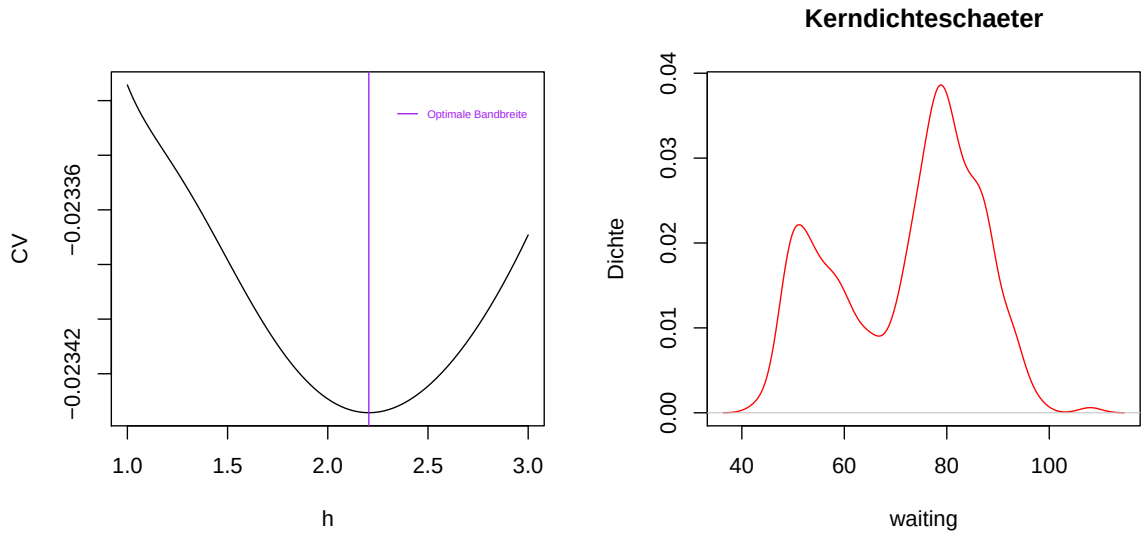

Abbildung 2.13: Verlauf von $LSCV(h)$ auf $h = [1,3]$ (Grafik links) und Kerndichteschätzer mit Glättungsparameter $\hat{h}_{lscv} = 2,2047$ (Grafik rechts) in Beispiel 2.3.1

und m_1, m_2 und m_{12} die entsprechenden Mächtigkeiten, dann gilt:

$$LSCV(k) = \frac{(k_1k_2)^{1/2}}{2\pi n}\{\frac{1}{2} + \frac{n-1}{32} + \frac{1}{n}\sum_{(i,j)\in(M_1\setminus M_{12})}\left(-4\left(Z_{ij2}^{k_2} - \frac{1}{8}\right)^2\right) + \frac{1}{n}\sum_{(i,j)\in(M_2\setminus M_{12})}\left(-4\left(Z_{ij2}^{k_1} - \frac{1}{8}\right)^2\right) - \frac{49m_{12}}{16n} + \frac{1}{n}\sum_{i=2,\ldots,n,j<i}^{(i,j)\notin(M_1\cup M_2)}\left(-4\left(\prod_{l=1}^{2} Z_{ijl}^{k_l} - \frac{1}{8}\right)^2\right)\} \tag{2.34}$$

Im Folgenden wird in zwei Fällen ($k_1, k_2 \to \infty$, und $k_1 \to \infty$) diskutiert, wie diskretisierte Daten die Zielfunktion $LSCV(k)$ beeinflussen. Die Schlussfolgerung im Fall $k_1 \to \infty$ gilt in analoger Weise auch für den Fall $k_2 \to \infty$.

- Für $k_1, k_2 \to \infty$:

$$LSCV(k) = \frac{(k_1k_2)^{1/2}}{2\pi n}\left(\frac{1}{2} - \frac{3m_{12}}{n}\right) \tag{2.35}$$

  $LSCV(k)$ geht gegen minus unendlich, wenn $m_{12} > n/6$;

- Für $k_1 \to \infty$:

$$LSCV(k) = \frac{(k_1k_2)^{1/2}}{2\pi n}\left(\frac{1}{2} + \frac{m_1}{16} - \frac{49m_{12}}{16} + \sum_{(i,j)\in(M_1\setminus M_{12})}\left(-4\left(Z_{ij2}^{k_2} - \frac{1}{8}\right)^2\right)\right) \tag{2.36}$$

Der Zielfunktionswert hängt von Z_{ij2} und k_2 folgendermaßen ab:

- $LSCV(k)$ hat das Supremum

$$LSCV(k)_{sup} = \frac{(k_1 k_2)^{1/2}}{2\pi n}\left(\frac{1}{2} + \frac{m_1 - 49m_{12}}{16n}\right) \tag{2.37}$$

  wenn $Z_{ij2}^{k_2} \to 1/8$, also $k_2(X_{2i} - X_{2j})^2 \to 4\log 8$, für $(i,j) \in (M_1 \backslash M_{12})$;
- $LSCV(k)$ hat das Infimum

$$LSCV(k)_{inf} = \frac{(k_1 k_2)^{1/2}}{2\pi n}\left(\frac{1}{2} - \frac{3m_1}{n}\right) \tag{2.38}$$

  wenn $Z_{ij2}^{k_2} \to 1$, also $k_2(X_{i2} - X_{j2})^2 \to 0$ für $(i,j) \in (M_1 \backslash M_{12})$.

Im bivariaten Fall haben M_{12} einen großen Einfluss auf $LSCV(k)$. $LSCV(k)$ geht gegen minus unendlich, wenn $m_{12} > n/6$. M_1 wirkt auf $LSCV(k)$ zusammen mit Bindungen bei der anderen Variable: falls sie fast gleich sind, dann geht $LSCV(k)$ gegen minus unendlich, wenn $m_1 > n/6$, aber falls es gilt $k_2(X_{2i} - X_{2j})^2 \to 4\log 8$ für $(i,j) \in (M_1 \backslash M_{12})$, dann wird $LSCV(k)$ gegen minus unendlich gehen, wenn $49m_{12} - m_1 > 8n$. Es ist problematisch, $\hat{h}_{lscv}$ beim **geyser** Datensatz zu bestimmen, weil $m_1 = 1095$ für Variable *waiting*, $m_2 = 1835$ für Variable *duration* und $m_{12} = 76 > n/6 = 49,833$. Es ist aber noch zu untersuchen, ob man im multivariaten Fall auch ein anderes Minimum für $\hat{h}_{lscv}$ anstatt der trivialen Lösung $\hat{h}_{lscv} = 0$ wie im univariaten Fall finden kann.

2.3.1.3 $LSCV(k)$ bei Daten mit mehr als 2 diskretisierten Variablen

Eine allgemeine Form von $LSCV(k)$ kann man wie folgt darstellen. Seien M_{v_a} die Indexmenge der Bindungen mit $a \in \{1,...,d\}$ und $v_a \in P_a$, P_a die Menge aller möglichen Permutationen der a Elemente, und m_{v_a} die entsprechende Mächtigkeit, dann gilt:

$$\begin{aligned} LSCV(k) = \frac{(k_1 k_2)^{1/2}}{2\pi n}\{\frac{1}{2^{d/2}} + \frac{n-1}{2^{d+3}} + \frac{1}{n}\{\sum_{g\in P_1}\sum_{(i,j)\in M_{v_1}}\left(-4\left(\prod_{l=1,l\neq g}^{d} Z_{ijl}^{k_l} - \frac{1}{4\cdot 2^{d/2}}\right)^2\right) \\ - \sum_{a=2,...,(d-1)}\sum_{g\in P_a}\sum_{(i,j)\in M_{v_a}}\left(-4\left(\prod_{l\in(1,...,d)}^{l\neq g} Z_{ijl}^{k_l} - \frac{1}{4\cdot 2^{d/2}}\right)^2\right)\} - \frac{m_{v_d}}{n}\left(-4\left(1 - \frac{1}{4\cdot 2^{d/2}}\right)^2\right) \\ + \frac{1}{n}\sum_{(i,j)\notin(\bigcup_{l=1}^{d}(M_l))}\left(-4\left(\prod_{l=1}^{d} Z_{lij}^{k_l} - \frac{1}{4\cdot 2^{d/2}}\right)^2\right)\} \end{aligned} \tag{2.39}$$

wobei $i = 2,...,n,\ j < i$. Man sieht in obiger Formel, dass ein allgemeines Kriterium nicht zu definieren ist, um festzustellen, in welcher Beziehung m_{v_a} zu n steht, so dass $LSCV(k)$ im Fall $k \to \infty$ also h sehr klein, gegen minus unendlich geht.

2.3.1.4 Bemerkung

Es ist in der Praxis üblich, dass die zu analysierenden Daten Bindungen enthalten. Zwei Hauptgründe sind:

- Die Bindungen kommen bei der Messung natürlicherweise vor;
- Die Daten sind in gewissem Masse diskretisiert worden, um den Speicherplatz zu ersparen oder die Rechenzeit für die Datenanalyse zu reduzieren.

Im Betrachten der guten Qualität von $\hat{h}_{lscv}$ ist es von Bedeutung, die LSCV Methode bei Daten mit Bindungen richtig anzuwenden. Eine einfach vorstellbare Lösung zu diesem Problem ist durch Hinzufügung gleichverteilter Störungsterme (Zychaluk & Patil 2006). Im multivariaten Fall ist es aber in manchen Fällen schwer zu entscheiden, zu welcher Variable die Störungsterme hinzu zu fügen sind, z.B., in **geyser** Daten enthalten beide Variablen Bindungen mit $m_1 = 1095$ für Variable *waiting* und $m_2 = 1835$ für Variable *duration*, aber die Diskretisierung von *waiting* hat eigentlich keinen großen Einfluss auf die Anwendung der LSCV Methode. Noch zu erwähnen ist, dass der Verlauf von $LSCV(h)$ vor der Datenmanipulation sorgfältig überprüft werden soll, weil ein globales Minimum an Null **nicht** bedeutet, dass man keine sinnvolle $\hat{h}_{lscv}$ finden kann.

2.3.2 Identifizierung eines geeigneten Minimums bei der BCV Methode

Beim Anwenden der BCV Methode kommen im gesuchten Bereich oft mehrere lokale Minima vor. Für die Auswahl eines geeigneten Minimums wird der folgende Vorschlag weit akzeptiert (Scott 1992):

> ... therefore, $\hat{h}_{BCV}$ is taken to be the largest local minimizer of $BCV(h)$ less than or equal to the oversmoothed bandwidth.

Im Folgenden werden die folgenden Punkte über dieses Problem anhand der in Abschnitt 2.2 benutzten Daten gezeigt:

1. Dieses Problem kommt oft vor;
2. Die obere Grenze (Oversmoothed Bandbreite) spielt dabei eine wichtige Rolle;
3. In der Regel soll man **nicht** immer das größte lokale Minimum nehmen.

Zuerst wird das Problem im univariaten Fall anhand der univariaten Daten aus Abschnitt 2.2 veranschaulicht. Abbildung 2.14-2.17 zeigen $BCV1(h)$ und $BCV2(h)$ auf $[0, 1h_{os}; 2h_{os}]$ bei 16 per Zufall ausgewählten Stichproben aus den 200 Stichproben aus $S1$ und $S2$ (vgl. Abs. 2.2), wobei $h_{os} = 1,144sn^{-1/5}$ die oversmoothed Bandbreite mit empirischer Standardabweichung s ist. Um (2) und (3) deutlich zu zeigen, markiert man h_{MISE} und h_{os} in den Grafiken mit lila und roter Linie. In Abbildung 2.14-2.17 sieht man folgendes:

1. $BCV1(h)$ und $BCV2(h)$ haben ein unique lokales Minimum bei Stichproben aus $S2$, das h_{MISE} entspricht;

2. $BCV2(h)$ hat zwei lokale Minima bei Stichproben aus S1. Im Intervall $[0, 1h_{os}; h_{os}]$ liegt das kleinere lokale Minimum, das h_{MISE} entspricht;

3. $BCV1(h)$ hat zwei lokale Minima im Intervall $[0, 1h_{os}; h_{os}]$ bei Stichproben aus S1. Das kleinere lokale Minimum spiegelt h_{MISE} wider.

Aus diesem Beispiel ist der folgende Schluss zu ziehen:

- Ob $BCV(h)$ in $[0, 1h_{os}; h_{os}]$ mehrere lokale Minima hat oder nicht, hängt von der unbekannten wahren Dichte ab;

- h_{os} wird als die obere Grenze des gesuchten Intervalls eingesetzt, um die evtl. Überglättung durch einen Kerndichteschätzer mit einem großen Glättungsparameter zu vermeiden. In Abbildung 2.15 sieht man, dass sich bei fast allen Grafiken ein größeres lokales Minimum in der Nähe von h_{os} befindet;

- In Abbildung 2.14-2.17 sieht man, dass die MISE-optimale Bandbreite durch das kleinere lokale Minimum von $BCV1(h)$ bzw. $BCV2(h)$ widerspiegelt worden ist, was dem obigen Vorschlag von Scott (1992) widerspricht.

In der explorativen Datenanalyse wird Kerndichteschätzung oft für die Identifizierung der Modalstruktur in den Daten verwendet. In diesen Fällen hat die richtige Auswahl der Bandbreite einen großen Einfluss auf das Resultat. Zur Veranschaulichung nimmt man $BCV1(h)$ bei Stichprobe 31 aus $S1$ als Beispiel. Dabei hat $BCV1(h)$ zwei lokale Minima $\hat{h}_1 = 0,0015$ und $\hat{h}_2 = 0,0033$ in $[0, 1h_{os}; h_{os}]$. Abbildung 2.3.2 zeigt die wahre Dichtefunktion von $S1$, die zwei Kerndichteschätzer mit Glättungsparameter $\hat{h}_1$ und $\hat{h}_2$. In Abbildung 2.3.2 sieht man, dass man mit $\hat{h}_2$ die wahre Dichte überschätzt und deswegen manche wichtige Eigenschaften der Daten verpasst, während die wahre Modalstruktur durch Dichteschätzer mit $\hat{h}_1$ gut widerspiegelt wird. Zum Zweck der Datenanalyse kann man natürlich verschiedene Glättungsparameter verwenden, um unterschiedliche Versionen der Daten zu bekommen, z.B., man bekommt in diesem Fall mit $\hat{h}_2$ einen groben Überblick über die Datenstruktur. Es stellt sich nun aber die Frage, welches lokale Minimum sollte man bei der BCV Methode nehmen im Sinne von MISE, wenn sich mehrere lokale Minima in $[0, 1h_{os}; h_{os}]$ befinden? Meine Empfehlung wäre, dass man das kleinste lokale Minimum nimmt.

Im multivariaten Fall hat man eben das Problem beim Identifizieren eines geeigneten lokalen Minimums beim Anwenden der BCV Methode. Mit der Erhöhung der Dimension wird es wegen grafischer Darstellung und des Rechenaufwands ziemlich schwierig, einen Überblick über das Verhalten von $BCV(h)$ zu bekommen. Im Folgenden wird das Problem an den bivariaten Daten aus Abschnitt 2.2 veranschaulicht.

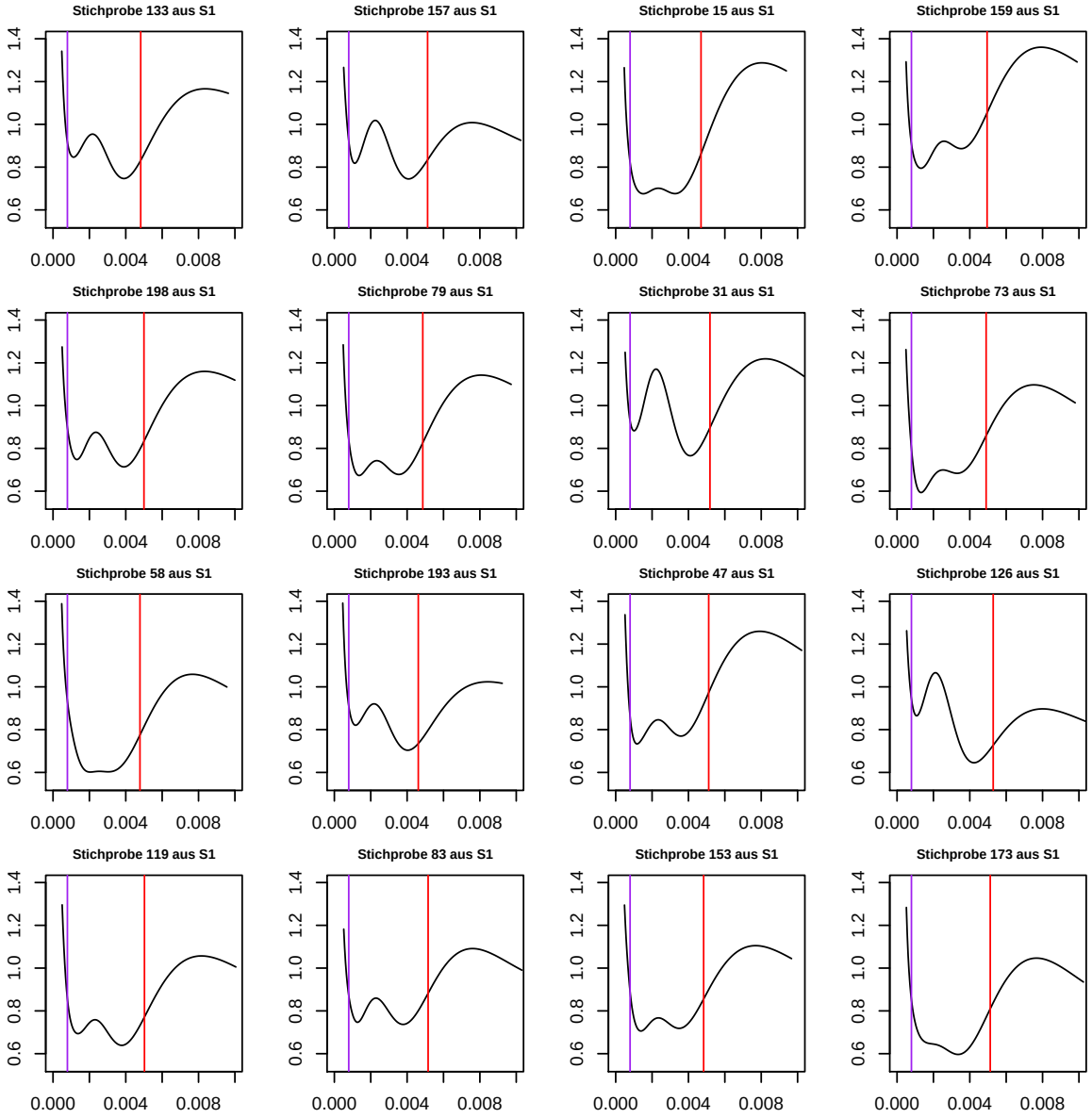

Abbildung 2.14: Verlauf von $BCV1(h)$ bei 16 per Zufall ausgewählten Stichproben aus den 200 Stichproben aus $S1$ in Abschnitt 2.2. h_{MISE} und h_{os} werden mit lila und roter Linie markiert.

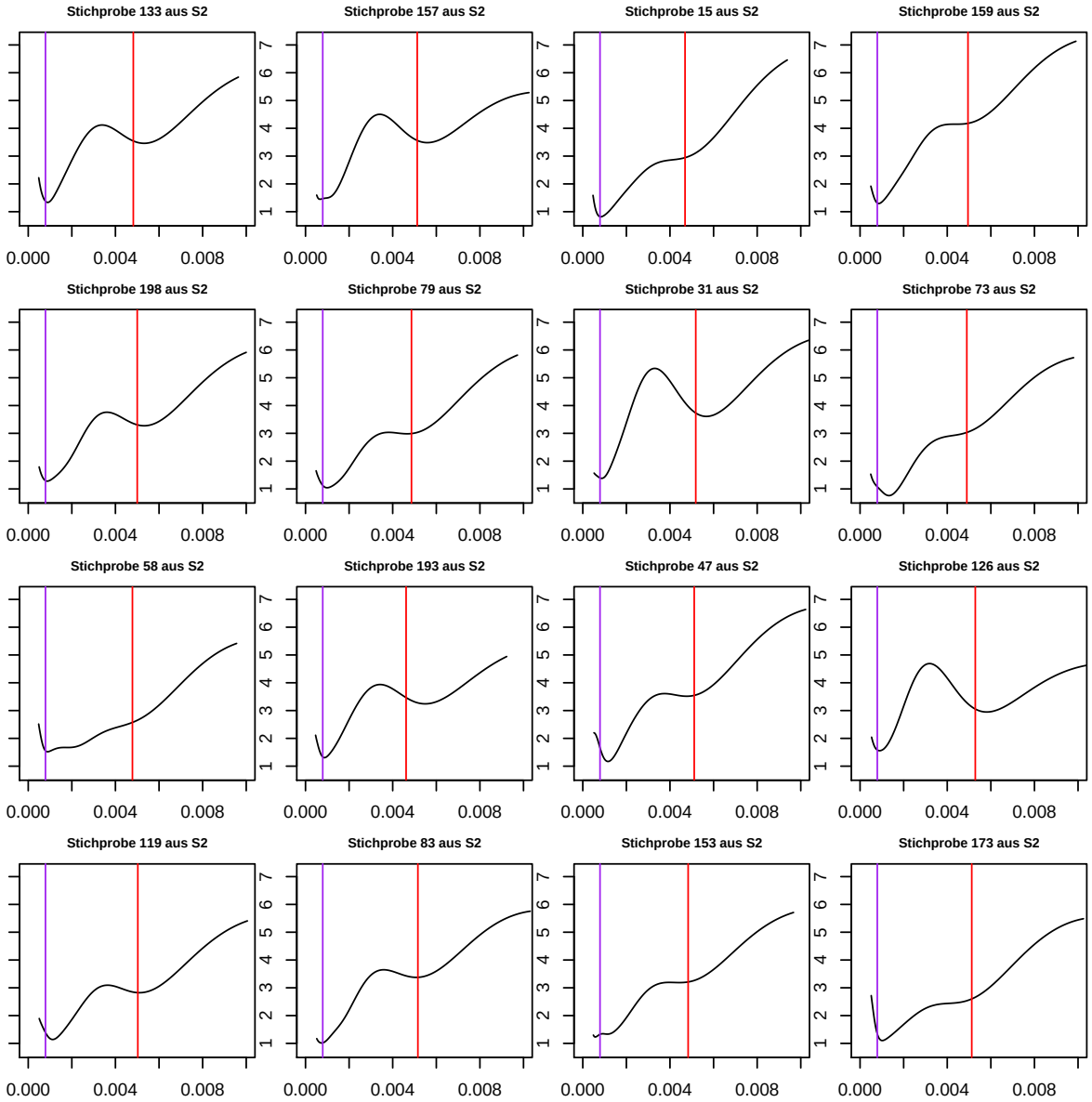

Abbildung 2.15: Verlauf von $BCV2(h)$ bei 16 per Zufall ausgewählten Stichproben aus den 200 Stichproben aus $S1$ in Abschnitt 2.2. h_{MISE} und h_{os} werden mit lila und roter Linie markiert.

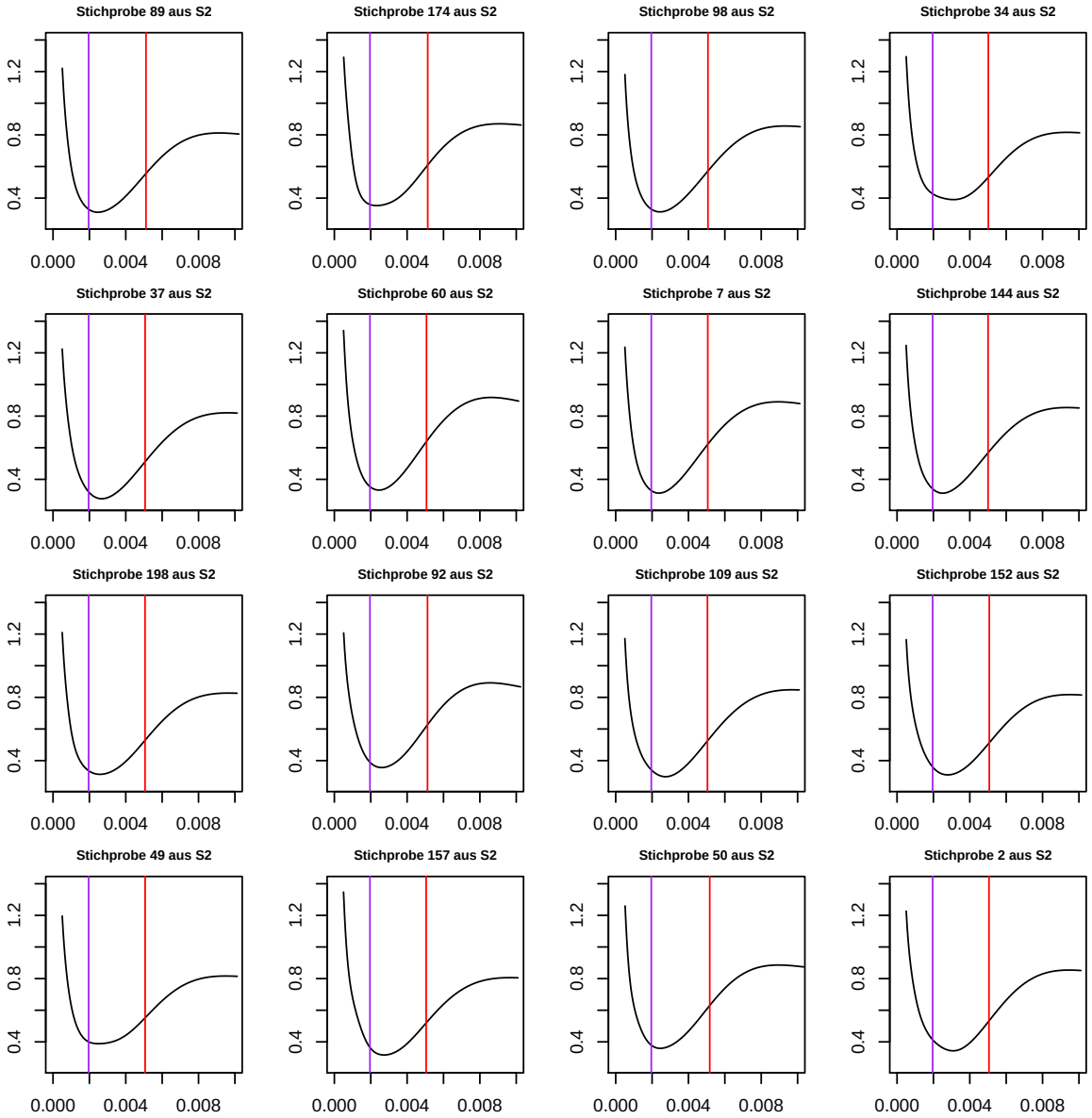

Abbildung 2.16: Verlauf von $BCV1(h)$ bei 16 per Zufall ausgewählten Stichproben aus den 200 Stichproben aus $S2$ in Abschnitt 2.2. h_{MISE} und h_{os} werden mit lila und roter Linie markiert.

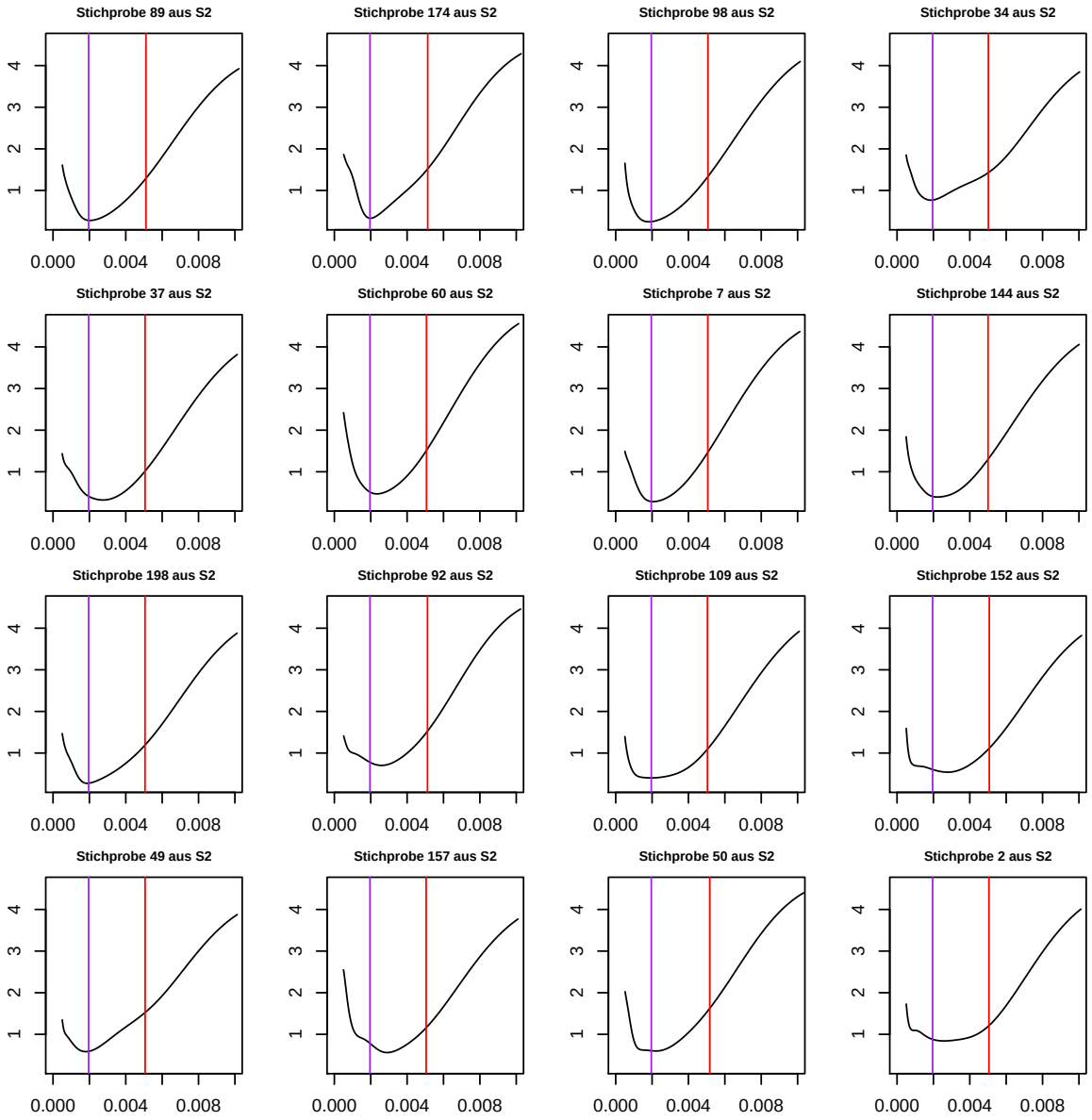

Abbildung 2.17: Verlauf von $BCV2(h)$ bei 16 per Zufall ausgewählten Stichproben aus den 200 Stichproben aus $S2$ in Abschnitt 2.2. h_{MISE} und h_{os} werden mit lila und roter Linie markiert.

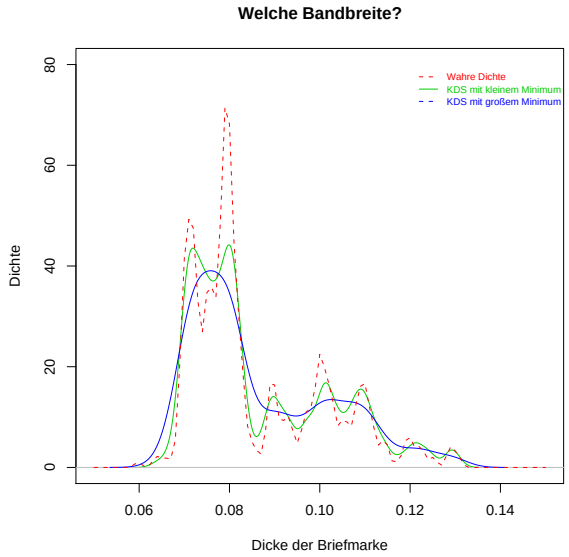

Abbildung 2.18: Wahre Dichtefunktion von $S1$ und zwei Kerndichteschätzer mit Glättungsparameter $\hat{h}_1 = 0,0015$ und $\hat{h}_2 = 0,0033$

Analog wie im univariaten Fall zieht man zufällig 9 Stichproben jeweils aus den 200 Stichproben mit Dichtefunktion f_{2d1} bzw. f_{2d2} und stellt bei denen das Verhalten von $BCV1(h)$ und $BCV2(h)$ auf $[0,1 \cdot h_{os}; 1,2 \cdot h_{os}]$ in Abbildung 2.19-2.22 im Imageplot dar. Bei grafischer Darstellung werden die Funktionswerte an 31×31 Gitterpunkten berechnet und deren Inverse in **R** Farbpalette cm.colors(10) abgebildet, wobei die großen (kleinen) Werte von $BCV1(h)$ und $BCV2(h)$ blau (rosa) eingefärbt sind. Als Referenz wird im Imageplot die MISE-optimale Bandbreite mit grünem Kreuz markiert.

In Abbildung 2.19-2.22 sieht man folgendes:

1. $BCV2(h)$ zeigt ein besseres Verhalten in dem Sinne, dass sie im gesuchten Bereich ein unique lokales Minimum hat;

2. Die kleinen Werte von $BCV1(h)$ befinden sich in einem großen zusammenhängenden Bereich;

3. $BCV1(h)$ ($BCV2(h)$) nimmt bei Stichprobe Nr. 62 und Nr. 97 (Nr. 148) mit Dichtefunktion f_{2d1} (f_{2d2}) ihre kleinen Werte in getrennten oder fast getrennten Bereichen, was impliziert, dass man verschiedene Lösungen unter unterschiedlichen Startvektoren bekommen kann.

Zur Veranschaulichung von 2 und 3 nimmt man die Bestimmung von $\hat{h}_{bcv1}$ bei Stichprobe Nr. 97 mit Dichtefunktion f_{2d1} und bei Stichprobe Nr. 50 mit Dichtefunktion f_{2d2} als Beispiel. In Abbildung 2.23 stellt man das Minimum-Suchen beim numerischen Verfahren

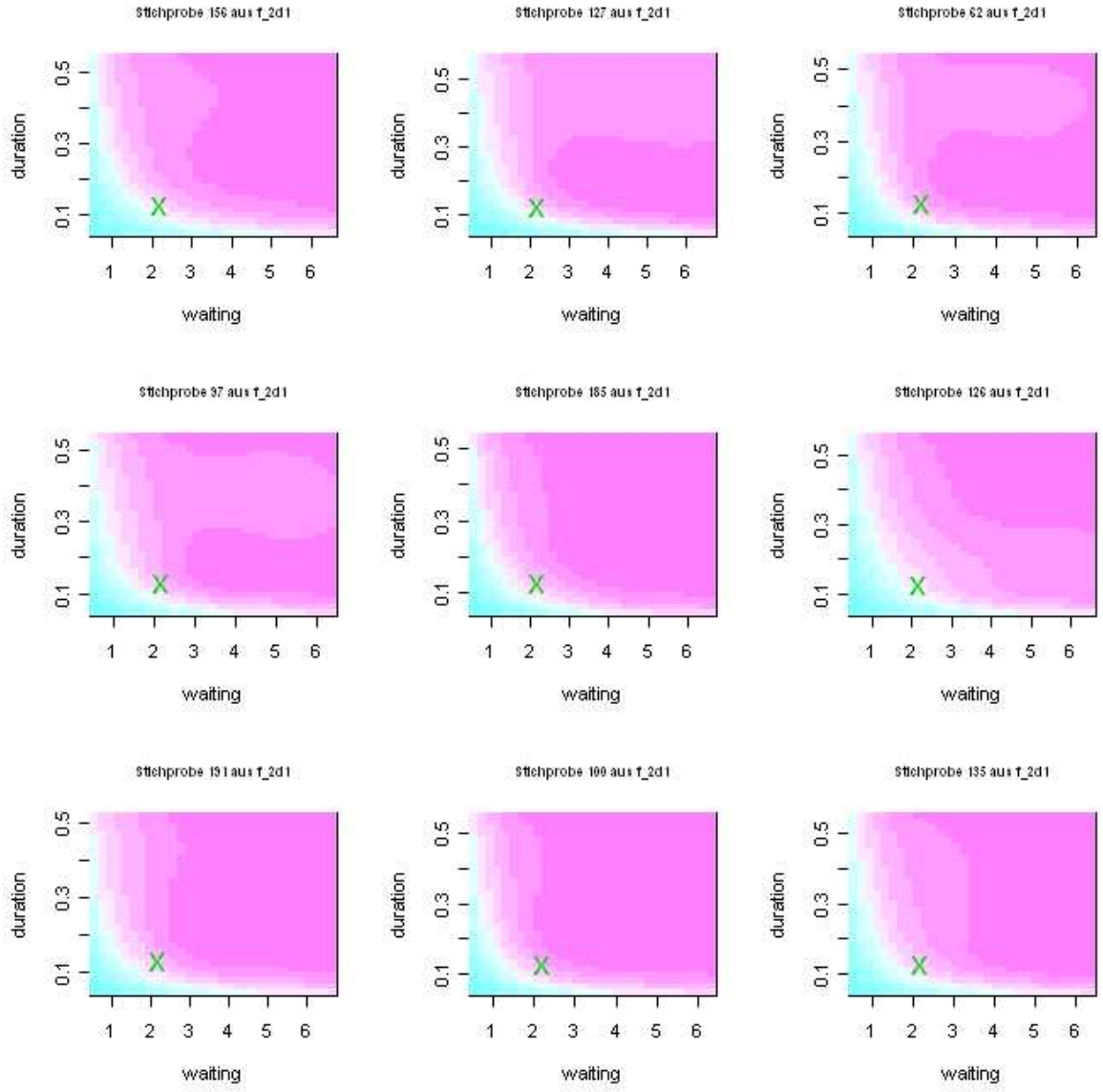

Abbildung 2.19: $BCV1(h)$ bei 9 per Zufall ausgewählten Stichproben aus den 200 Stichproben mit Dichtefunktion f_{2d1} in Abschnitt 2.2. h_{MISE} in grünem Kreuz.

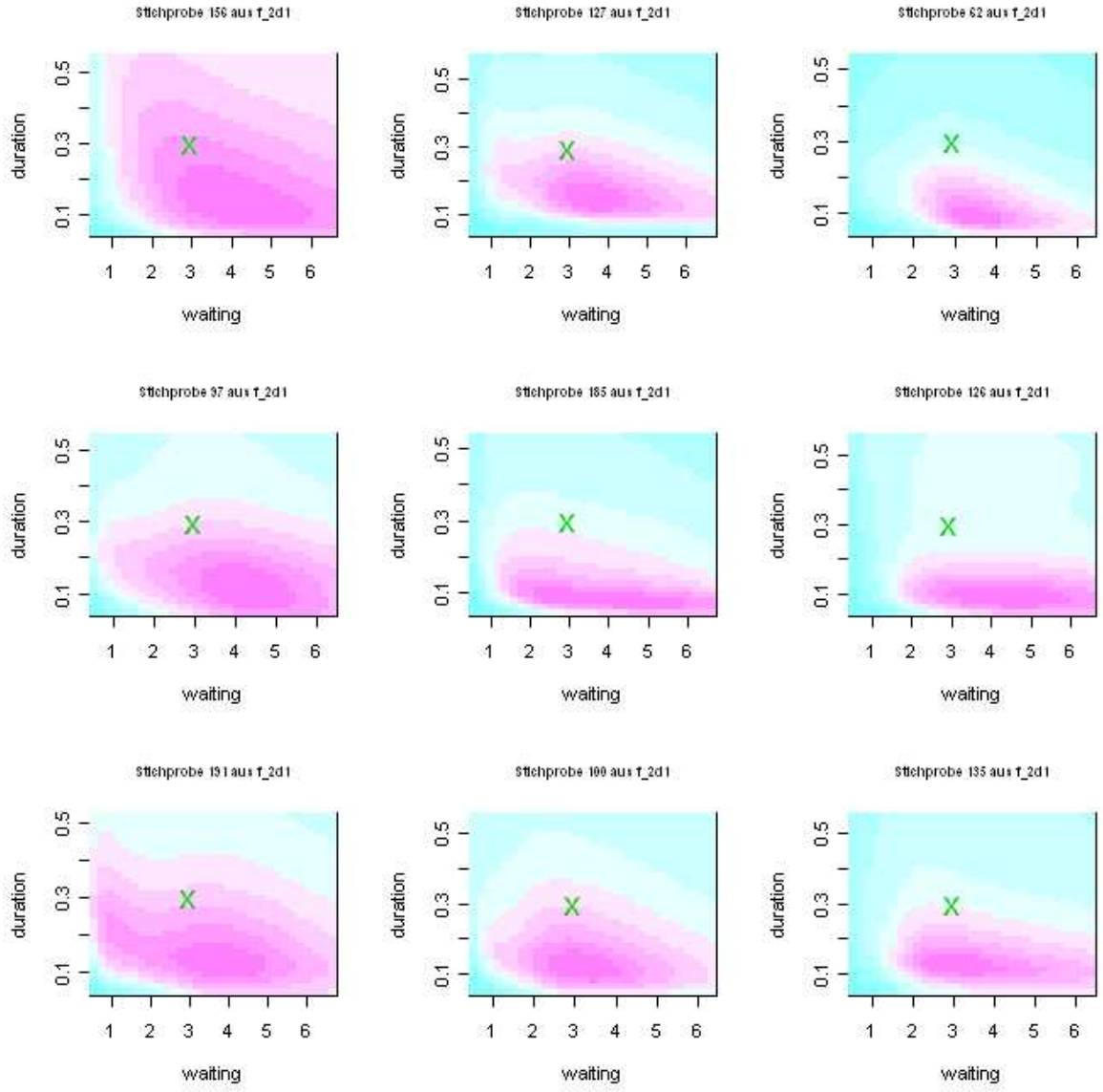

Abbildung 2.20: $BCV2(h)$ bei 9 per Zufall ausgewählten Stichproben aus den 200 Stichproben mit Dichtefunktion f_{2d1} in Abschnitt 2.2. h_{MISE} in grünem Kreuz.

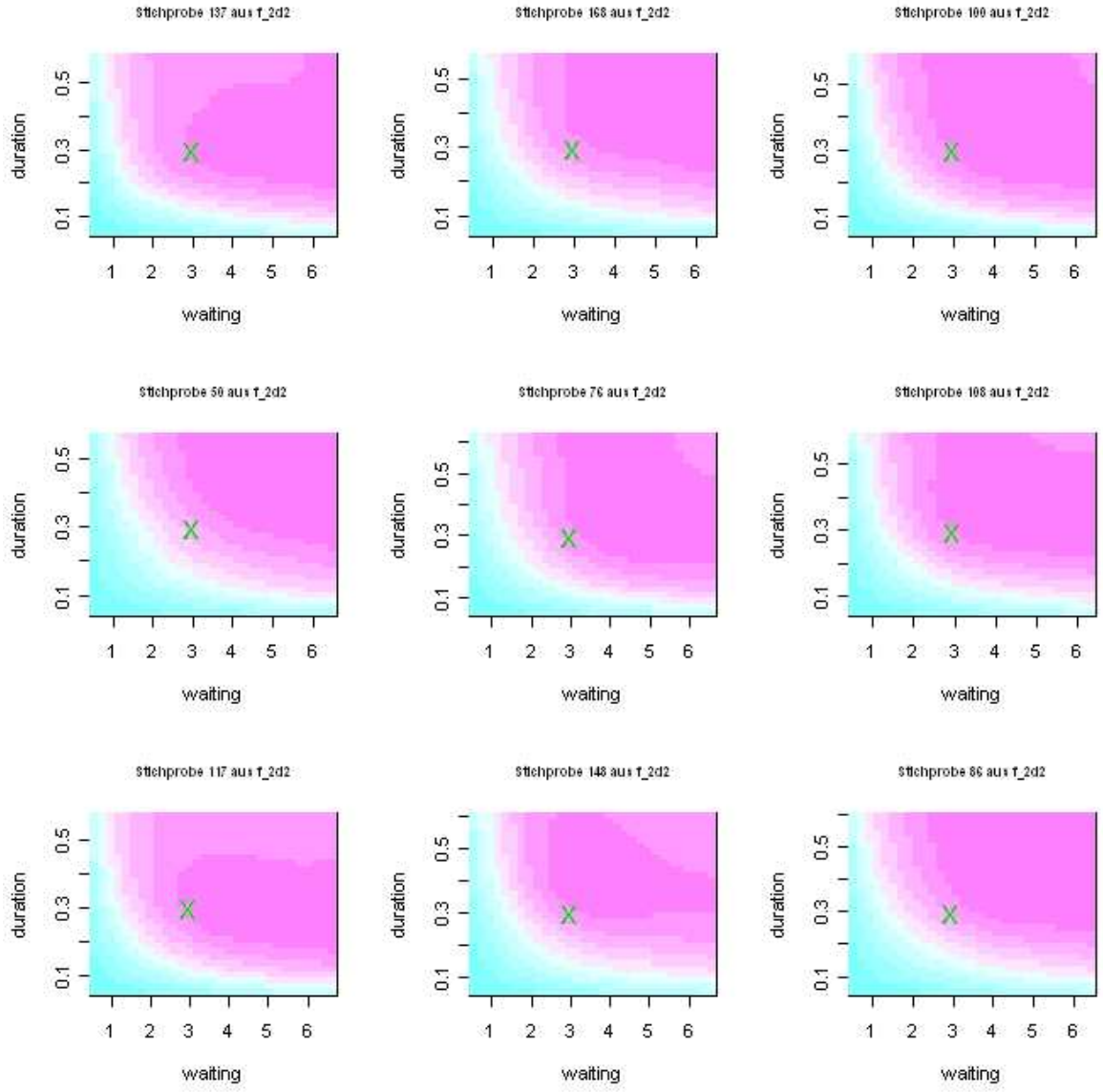

Abbildung 2.21: $BCV1(h)$ bei 9 per Zufall ausgewählten Stichproben aus den 200 Stichproben mit Dichtefunktion f_{2d2} in Abschnitt 2.2. h_{MISE} in grünem Kreuz.

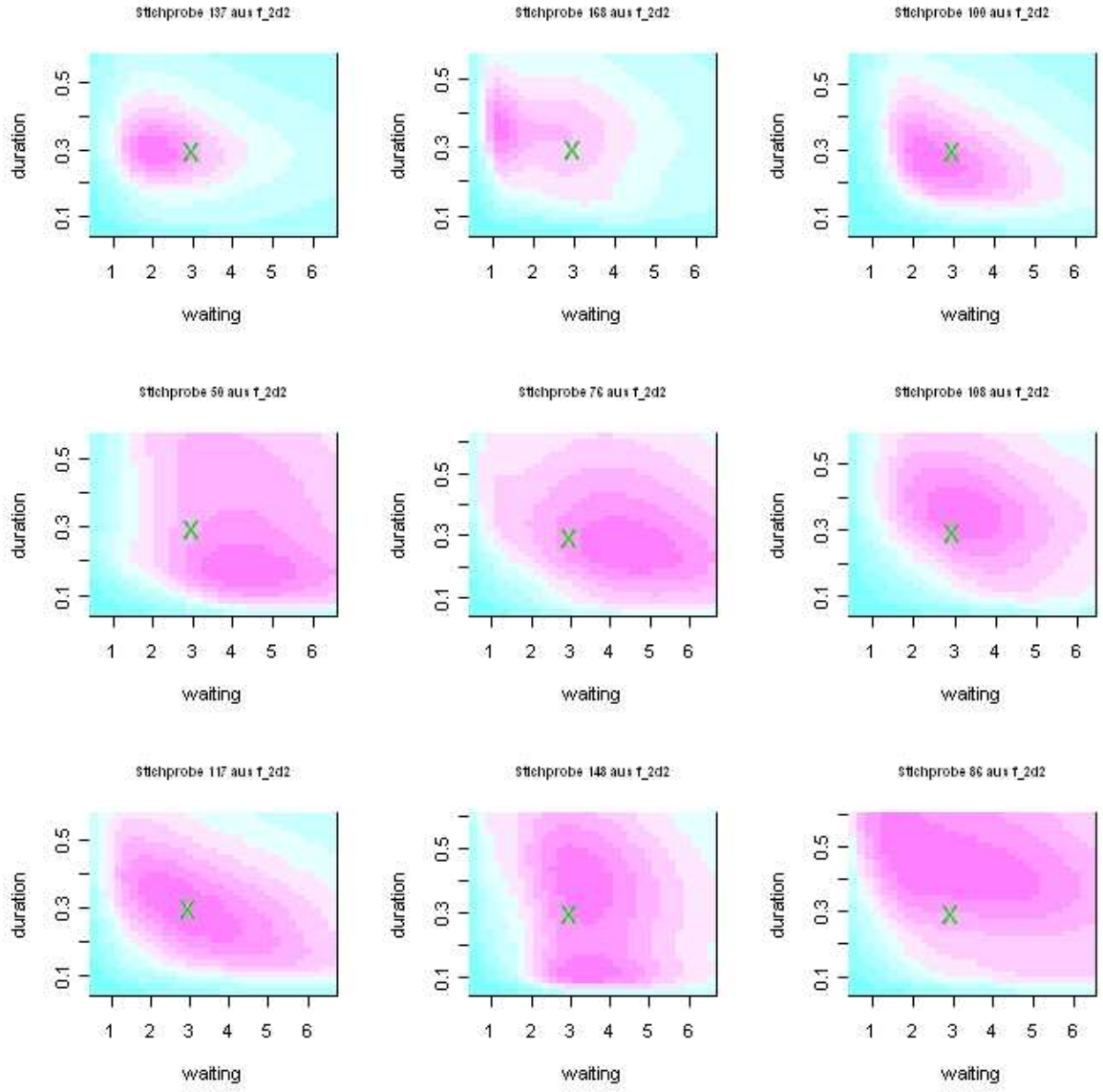

Abbildung 2.22: $BCV2(h)$ bei 9 per Zufall ausgewählten Stichproben aus den 200 Stichproben mit Dichtefunktion f_{2d2} in Abschnitt 2.2. h_{MISE} in grünem Kreuz.

grafisch dar, wobei $(0,2;0,2)^T.*h_{os}$, $(0,2;1)^T.*h_{os}$, $(1;0,2)^T.*h_{os}$ und $(1;1).*h_{os}$ (.* steht für komponentenweises Produkt) jeweils als Startvektoren ausgewählt werden. In der

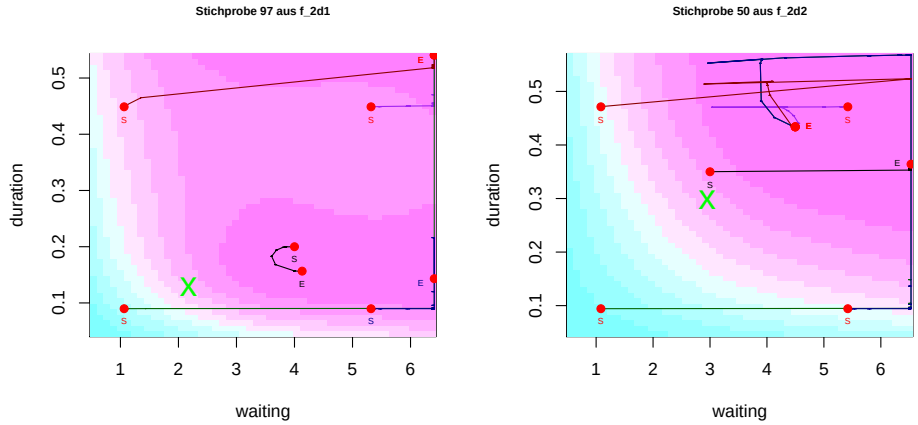

Abbildung 2.23: Minimum-Suchen beim numerischen Verfahren bei Stichprobe Nr. 97 (50) mit Dichtefunktion f_{2d1} (f_{2d2})

Grafik in Abbildung 2.23 steht die dunkle grüne/dunkle rote/dunkle blaue/lila Linie für den Suchweg mit Startvektor $(0,2;0,2)^T.*h_{os}/(0,2;1)^T.*h_{os}/(1;0,2)^T.*h_{os}/(1;1)^T.*h_{os}$. Die Startposition und Endeposition des Suchwegs werden in der Grafik in S und E markiert. S und E werden gleich eingefärbt, falls die entsprechenden Start- und Endepunkte aus dem selben Suchweg sind. Wenn ein Endepunkt mehrere Startpunkte hat, dann werden sie alle gleich eingefärbt. In Abbildung 2.23 sieht man folgendes:

- Das Suchverfahren mit den obigen vier Startvektoren liefert bei Stichprobe Nr. 97 mit Dichtefunktion f_{2d1} kein gutes Resultat, während man mit einem anderen Startpunkt (hier $(4;0,2)^T$ als Beispiel) aus dem rosa Bereich unten rechts eine sinnvollere Bandbreite bekommt;
- Das Suchverfahren mit den vier Startvektoren liefert bei Stichprobe Nr. 50 mit Dichtefunktion f_{2d2} ein gutes Resultat, während man mit einem anderen Startpunkt (hier $(3;0,35)^T$ als Beispiel) aus dem Mittelbereich ein verzerrtes Ergebnis bekommt.

Abbildung 2.24 zeigt beispielhaft die zwei Dichteschätzer für f_{2d1} anhand der Stichprobe Nr. 97 mit Glättungsparametern aus numerischen Verfahren mit Startpunkten h_{os} (Grafik links) und $(4;0,2)^T$ (Grafik rechts). In Abbildung 2.24 ist deutlich zu erkennen, dass der Kerndichteschätzer in der linken Grafik zur Überglättung führt. Generell wird ein großer Bias entstehen, wenn man das Problem der oben erwähnten Minimum-Identifizierung beim Anwenden der BCV Methode nicht beachtet. Ein möglicher Lösungsvorschlag ist, mehrere Startvektoren beim numerischen Optimierungsverfahren zu benutzen, um alle lokale Minima im gesuchten Gebiet finden zu können. In der

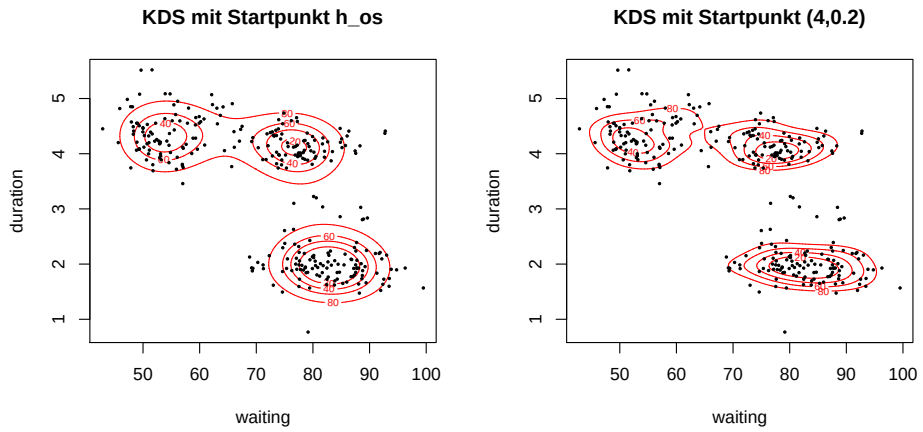

Abbildung 2.24: Zwei Kerndichteschätzer für f_{2d1} anhand der Stichprobe Nr. 97 mit Bandbreiten aus numerischen Verfahren mit Startpunkt h_{os} (Grafik links) und $(4; 0, 2)^T$ (Grafik rechts)

explorativen Datenanalyse sollte man ein sinnvolles Minimum oder ein paar geeignete Minima für Kerndichteschätzung auswählen, um die richtige Version oder verschiedene wichtige Versionen von den Daten zu gewinnen. Im Unterschied zur Literatur bevorzuge ich genauso wie im univariaten Fall das kleinere lokale Minimum, falls sich mehrere lokale Minima im gesuchten Bereich befinden, weil die Daten in diesem Fall mehr lokale Eigenschaften haben und ein Kerndichteschätzer mit kleinerer Bandbreite die lokalen Eigenschaften der Daten gut widerspiegeln kann. Dies ist aber noch weiter zu untersuchen.

2.4 Zusammenfassung

In diesem Kapitel werden vier Bandbreitenselektoren (LCV, LSCV, BCV (BCV1 und BCV2), DPI) anhand simulierter Daten verglichen. Das Ziel dieses Vergleichs besteht darin, das Verständnis der Rolle des Glättungsparameters bei Kerndichteschätzung zu vermitteln. Es hat sich ergeben, dass die LSCV Methode im Betrachten von MISE-optimaler Bandbreite die beste Leistung bringt. Eigentlich haben die klassischen Methoden (LCV, LSCV) in unseren Beispielen ihren Vorteil gezeigt, nämlich die volle Orientierbarkeit an den empirischen Daten. Aufgrund dieser Eigenschaft sind die LCV und LSCV Methoden in der explorativen Datenanalyse zu empfehlen.

Es ist schon bekannt, dass die LSCV Methode beim Datensatz mit Bindungen nicht einwandfrei anwendbar ist. In der Literatur wird dieses Problem im univariaten Fall gut analysiert. In diesem Kapitel wurde die Untersuchung dieses Problems anhand einer Umformung von $LSCV(h)$ auf multivariaten Fall erweitert. Das Verhalten von

$LSCV(h)$, $h \to 0$ kann man im bivariaten Fall festlegen. Es wurde auch gezeigt, dass dieses Verhalten im mehrdimensionalen Fall kompliziert ist und von Fall zu Fall beschrieben werden soll. Zu beachten ist, dass man in manchen Fällen trotz des globalen Minimums von $LSCV(h)$ an $h = 0$ eine sinnvolle Bandbreite bekommen kann, was in der Literatur oft ignoriert worden ist.

In diesem Kapitel wurde das Problem der Auswahl eines geeigneten Minimums bei BCV anhand der Beispiele aus Abschnitt 2.2 erläutert und veranschaulicht. Im Unterschied zu dem Auswahlvorschlag von Scott (1992) wird in dieser Arbeit das kleinere lokale Minimum bevorzugt in den Fällen, in denen sich mehrere lokale Minima im gesuchten Gebiet befinden. Da die Minimum-Suche durch numerische Verfahren funktioniert und in der Regel vom Startvektor abhängt, spielt die Auswahl eines passenden Startvektors beim Bestimmen von $\hat{h}_{bcv}$ eine wichtige Rolle. In der Praxis der Datenanalyse ist zu empfehlen, dass man beim numerischen Verfahren verschiedene Startvektoren benutzen soll, um alle lokale Minima von $BCV(h)$ zu finden und die entsprechenden Bandbreiten bei Kerndichteschätzung zu verwenden, damit man keine nützliche Information der Daten verliert.

3 Dichteschätzung und Clusteranalyse

Ein Fest-Kerndichteschätzer kann die wahre unbekannte Dichte in manchen Fällen nicht gut widerspiegeln, insbesondere wenn die Daten einen langen Schwanz (Silverman (1986)) oder eine Multimodal-Struktur haben. Bei multimodalen Daten ist es in der Regel schwierig, eine geeignete feste Bandbreite für die richtige Schätzung von wahren Modi und den Regionen zwischen wahren Modi auszuwählen. Im hochdimensionalen Fall leidet die Fest-Kerndichteschätzung stark unter „Curse of Dimensionality“ , das Scott (1992) wie folgt

> ... the apparent paradox of 'neighborhoods' in higher dimensions - if the neighborhoods are 'local', then they are almost surely 'empty', whereas if a neighborhood is not 'empty', then it is not 'local'.

beschrieb, es sei denn, falls der Umfang der Daten riesig ist.

Um den Schätzungsfehler bei Kerndichteschätzung zu reduzieren, wurde seit der Geburt von Kerndichteschätzung eine Vielzahl von statistischen Methoden vorgeschlagen. Es wurde gezeigt, dass man den Bias bei Kerndichteschätzung durch Einbezug von adaptivem Glättungsparameter reduzieren kann (Terrell & Scott (1992), Sain (1994), Sain & Scott (1996), Sain (2002)). In Sain (2002) wurden zwei Algorithmen für die Konstruktion des Adaptiv-Kerndichteschätzers für die praktische Anwendung vorgestellt. Eigentlich ist ein Adaptiv-Kerndichteschätzer eine spezielle Form vom gemischten Modell, das in Scott & Szewczyk (2000) wie folgt definiert wurde,

$$\hat{f}(x,\theta) = \sum_{i=1}^{m} w_i f_i(x,\theta_i) \tag{3.1}$$

wobei $w_i > 0$ für $i = 1, ..., m$, $\sum_{i=1}^{m} w_i = 1$ und f_i eine beliebige Dichtefunktion mit Parameter θ_i. Ein bekanntes und weit verbreitetes gemischtes Modell ist wohl das Modell aus dem Model Based Clustering von Fraley & Raftery (2002).

Obwohl ein Fest-Kerndichteschätzer im multivariaten Fall an sich in der Regel kein guter Schätzer der unbekannten wahren Dichte ist, wird die multivariate Fest-Kerndichteschätzung in der explorativen Datenanalyse für die Aufdeckung der Modalstruktur (Mode Hunting) und das darauf basierte Clustering breit angewendet, wobei angenommen wird, dass die Daten einer gewissenWahrscheinlichkeitsverteilung unterliegen und zu den Domänen der Modi der entsprechenden Dichtefunktion zugeordnet werden können. Das Dichteschätzer basierte Clustering beruht auf den Grundideen von „Level Mode Analysis“ von Wishart (1969) und „High Density Cluster“ von Hartigan (1975). Typische Methoden sind z.B.,

DBSCAN von Ester et al. (1996), Generalized Single Linkage Methode von Stuetzle et al. (2007). Ein Fest-Kerndichteschätzer liefert unvermeidbar manche nicht signifikante Modi und deswegen ist es wichtig, die signifikanten Modi zu identifizieren und die Daten dementsprechend zuzuordnen. Gute Vorschläge für die Untersuchung der Modalstruktur in den Daten findet man in Minnotte & Scott (1993), Chaudhuri & Marron (1997), Godtliebsen et al. (2002), Duong et al. (2007) und Stuetzle et al. (2007).

Das vorliegende Kapitel gliedert sich in 4 Teile. Zunächst im ersten Teil vergleicht man 6 Varianten des Modells in (3.2) anhand eines simulierten Beispiels. Dann im zweiten Teil stellt man die statistischen Methoden vor, mit denen man die Modalstruktur bzw. Clusterstruktur der Daten auf Basis der Dichteschätzung untersuchen kann. Im dritten Teil wird das Dichteschätzer basierte hierarchische Verfahren anhand simulierter und praktischer Daten ausführlich erläutert. Im letzten Teil dieses Kapitels wird anhand eines praktischen Beispiels erläutert, wie man das Dichteschätzer basierte hierarchische Verfahren anwendet in dem Fall, wenn der Umfang der Daten groß ist.

3.1 Kerndichteschätzer und Gemischtes Modell

Seien $X_1, ..., X_n \in R^d$ eine Zufallsstichprobe aus einer Wahrscheinlichkeitsverteilung mit unbekannter Dichtefunktion f, dann wird f hier durch

$$\hat{f}(x, \theta) = \sum_{i=1}^{m} w_i \phi_i(x, \theta_i) \tag{3.2}$$

geschätzt, wobei $w_i > 0$ für $i = 1, ..., m$, $\sum_{i=1}^{m} w_i = 1$ und ϕ_i eine Normaldichte mit Parameter $\theta_i = (\mu_i, \Sigma_i)$ mit $\mu_i = (\mu_{i1}, \mu_{i2}, ..., \mu_{id})^T$ und Σ_i die Varianz-Kovarianz-Matrix. Man sieht in (3.2), dass die Anzahl der zu schätzenden Parameter von Modell zu Modell variiert. Die extremen Fälle sind: Modell mit 2 zu schätzenden Parametern, wobei $w_i = 1$, $\mu_1 = \mu_2 = ... = \mu_m = \mu 1_d$ (1_d ist der Vektor mit d Eins), $\Sigma_1 = \Sigma_2 = ... = \Sigma_m = h^2 I_d$ (I_d ist eine $d \times d$ Einheitsmatrix), und Modell mit $m(d+1)(d+2)/2$ zu schätzenden Parametern, wobei alle Parameter unterschiedlich sind. In diesem Abschnitt werden die folgenden 6 Varianten von dem Modell in (3.2) anhand eines simulierten Beispiels vorgestellt und graphisch veranschaulicht: Fest-Kerndichteschätzer mit diagonaler LSCV-Bandbreitenmatrix, zwei Binned Sample-Point Kerndichteschätzer von Sain (2002) mit diagonaler Bandbreitenmatrix, Gemischtes Modell aus dem Model Based Clustering von Fraley & Raftery (2002), Gemischtes Modell aus dem CEM-Algorithmus und Gemischtes Modell aus einem modifizierten SEM-Algorithmus von Celeux & Govaert (1992).

Beispiel 3.1.1: 200 simulierte Daten aus bivariater gemischten Normalverteilung mit 7 Komponenten
Die Daten werden wie folgt simuliert:

- Aus $(0,1; ...; 0,9) \times (0,1; ...; 0,9)$ 81 Punkten werden 7 Punkte per Zufall als Erwar-

tungswerte μ_i, $i = 1, ..., 7$ für die 7 Komponenten der gemischten Normalverteilung ausgewählt;

- Jede Komponente hat eine diagonale Varianz-Kovarianz-Matrix Σ_i, $i = 1, ..., 7$, deren Diagonalelemente per Zufall aus $(1/60, ..., 9/60) \times (1/60, ..., 9/60)$ ausgewählt werden;
- Das Gewicht der Komponenten w_i, $i = 1, ..., 7$ in der gemischten Normalverteilung ist proportional zu der Summe der Diagonalelemente von Σ_i;
- Die Daten aus der i-ten Komponente werden mit Wahrscheinlichkeit w_i gezogen.

In Beispiel 3.1.1 werden bivariate Daten simuliert, weil die Daten in der 2D Ebene grafisch darstellbar sind. Abbildung 3.1 zeigt die simulierten Daten, wobei man die Daten aus verschiedenen Komponenten unterschiedlich einfärbt und die Positionen der Erwartungswerte der 7 Komponenten mit „x“ markiert. Im Folgenden zeigt man die 6 Varian-

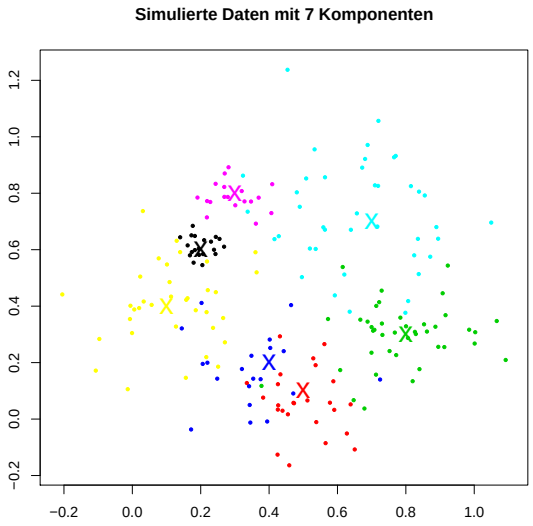

Abbildung 3.1: Simulierte Daten mit 7 Komponenten in Beispiel 3.1.1. Daten aus verschiedenen Komponenten unterschiedlich eingefärbt. μ_i, $i = 1, ..., 7$ mit „x“ markiert

ten des Modells in (3.2) anhand des Beispiels 3.1.1.

Modell 1 Fest-Kerndichteschätzer mit diagonaler LSCV-Bandbreitenmatrix

Dabei werden die Modellparameter in (3.2) wie folgt definiert: $m = n$, $w_i = 1/n$, $\mu_i = X_i$, $\Sigma = \Sigma_1 = \Sigma_2 = ... = \Sigma_n = argmin_\Sigma(LSCV)$, wobei Σ für eine diagonale Matrix steht. In Abbildung 3.2 wird **Modell 1** im Contour-Plot dargestellt. Man sieht in Abbil-

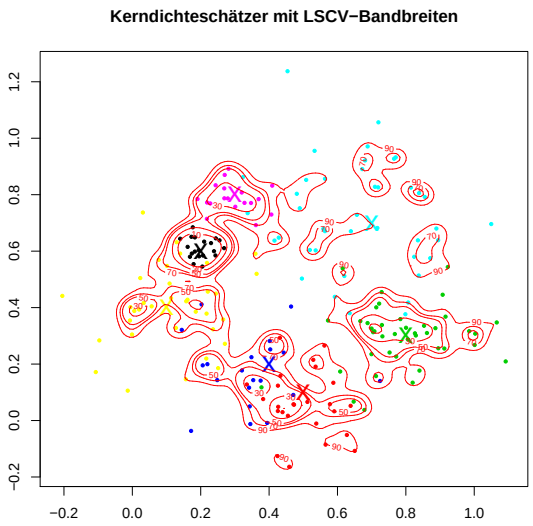

Abbildung 3.2: Fest-Kerndichteschätzer mit diagonaler LSCV-Bandbreitenmatrix im Contour-Plot in Beispiel 3.1.1

dung 3.2, dass drei Komponenten (grün/schwarz/rosa eingefärbt) dadurch identifiziert werden, während die anderen 4 Komponenten nicht nur einen sondern mehrere Modi des Dichteschätzers enthalten, was impliziert, dass der Fest-Kerndichteschätzer mit der LSCV-Bandbreitenmatrix eine Unterglättung darstellt.

Modell 2 Binned Sample-Point Kerndichteschätzer von Sain (2002) mit diagonaler Bandbreitenmatrix

Den Binned Sample-Point Kerndichteschätzer von Sain (2002) definiert man wie folgt:

$$\hat{f}_s(x) = \frac{1}{n} \sum_{j=1}^{m} n_j K_{H_j}(x - t_j) \tag{3.3}$$

wobei m die Anzahl von Bins, n_j die Anzahl von Daten im j-ten Bin, K die Kernfunktion, t_j der Mittelpunkt des j-ten Bins oder der Mittelpunkt der Daten im j-ten Bin, H_j die Bandbreitenmatrix für die Daten im j-ten Bin, und $w_j = n_j/n$, $\mu_j = t_j$, $\Sigma_j = H_j$ im Vergleich zu (3.2). Es gibt folgende zwei Varianten beim Konstruieren von $\hat{f}_s(x)$ in Sain (2002):

V1 Man teilt den Datenraum in Bins, deren Mittelpunkte als t_j genommen werden, und berechnet H_j nach einem gewissen Kriterium (hier LSCV), wobei die leeren Bins nicht berücksichtigt werden;

V2 Man nimmt zuerst einen Fest-Kerndichteschätzer als Pilot-Dichteschätzer und dann

ordnet die Daten zu dem am nächsten liegenden Modus des Pilot-Dichteschätzers zu. Man berechnet t_j und H_j diesem Clustering entsprechend.

Das LSCV-Kriterium für die Bestimmung von H_j bei Binned Sample-Point Kerndichteschätzung wird hier verwendet und sieht wie folgt aus (Sain (1999)):

$$LSCV = \int (\hat{f}_s(x))^2 dx + \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} n_{ij}^* K_{H_j}(X_i - t_j) \tag{3.4}$$

wobei $n_{ij}^* = n_j - 1$, falls $X_i \in B_j$, und $n_{ij}^* = n_j$, ansonsten, wobei B_j für den j-ten Bin steht.

Im Folgenden stell man **V1** und **V2** anhand des Beispiels 3.1.1 vor.

Zu **V1**: Man teilt den Datenraum in 7×7 Bins, davon gibt es 16 leere Bins und 9 Einzel-Punkt-Bins (d.h., diejenige Bins, die nur einen einzelnen Punkt enthalten).

Bemerkung:

H_j wird hier durch numerische Verfahren bestimmt. Mit der Erhöhung der Anzahl von Bins kann der Schätzungsfehler reduziert werden, dafür hat man aber mehr unbekannte Parameter beim numerischen Verfahren zu bestimmen. Als ein Kompromiss werden hier 49 Bins genommen.

Abbildung 3.3 zeigt die Daten in 49 Bins und die entsprechenden Gitterlinien. Beim

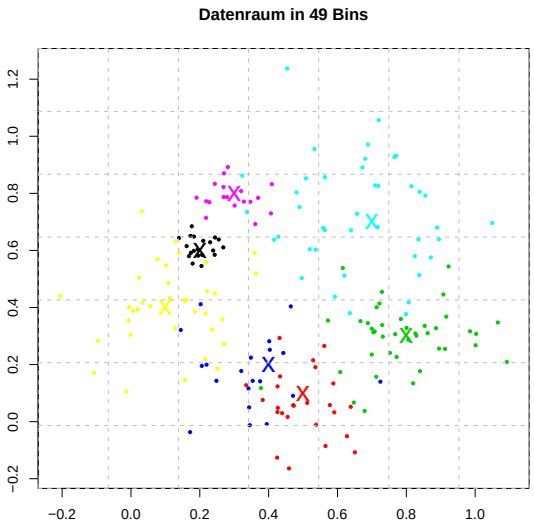

Abbildung 3.3: Daten in Beispiel 3.1.1 in 49 Bins und die entsprechenden Gitterlinien

numerischen Verfahren werden die 16 leeren Bins nicht berücksichtigt und die 9 Einzel-

Punkt-Bins mit dem Nachbarn kombiniert. Dadurch gibt es 27 Bins ($m = 27$ in (3.3)) mit 54 Diagonalelementen aus H_j, $j = 1, ..., 27$, die beim numerischen Verfahren nach dem LSCV-Kriterium in (3.4) zu schätzen sind. Abbildung 3.4 zeigt **V1** im Contour Plot. Man

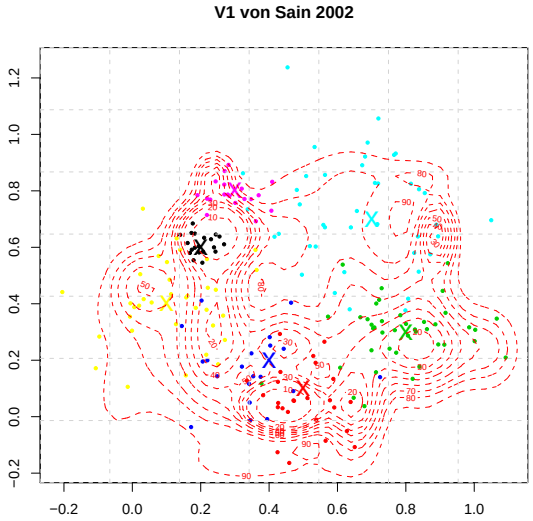

Abbildung 3.4: **V1** von Sain (2002) im Contour-Plot in Beispiel 3.1.1

sieht in Abbildung 3.4, dass **V1** zwar eine gemischte Normalverteilung darstellt aber die Formen der 7 Komponenten in der gemischten Normalverteilung nicht gut widerspiegeln kann. In der Tat ist der Dichteschätzer **V1** stark von der Auswahl der Anzahl und Position von Bins abhängig und deshalb aus den folgenden zwei Hauptgründen nicht zu empfehlen:

1. Falls die Anzahl von Bins gering ist, dann hat **V1** keinen Vorteil gegenüber einem Fest-Kerndichteschätzer in Bezug auf den Schätzungsfehler, und wenn man noch beachtet, dass in diesem Fall die Form eines Bins in **V1** kaum mit einer Komponente des gemischten Modells in Beispiel 3.1.1 übereinstimmt, dann ist **V1** eigentlich ein schlechter Schätzer;

2. Falls die Anzahl von Bins groß ist, dann hat man viele Parameter beim numerischen Verfahren zu schätzen. Es ist deswegen fraglich, ob das Resultat noch zuverlässig ist.

Zu **V2**: Man verwendet den Fest-Kerndichteschätzer aus **Modell 1** als Pilot-Dichteschätzer für die Konstruktion von **V2**. Mit dem Hill-Climbing Algorithmus (vgl. Abs. 1.3) werden 19 Modi im Pilot-Dichteschätzer gefunden. In Abbildung 3.5 zeigt man die Zuordnung der Daten zu den 19 Modi, indem man die Modi und die dazu zugehörigen Daten mit

Linien verbindet. Von den 19 Gruppen gibt es 5 Einzel-Punkt-Gruppen, die beim numerischen Verfahren für die Bestimmung von H_j zuerst mit dem Nachbarn kombiniert werden sollen. Beim numerischen Verfahren nimmt man die Position der Modi als t_j und die empirische Varianz von den zu t_j zugeordneteten Daten zu t_j als Anfangswert für H_j. H_j wird dann nach dem LSCV-Kriterium in (3.4) durch das numerische Verfahren bestimmt. In Abbildung 3.6 wird **V2** im Contour-Plot dargestellt. Man sieht

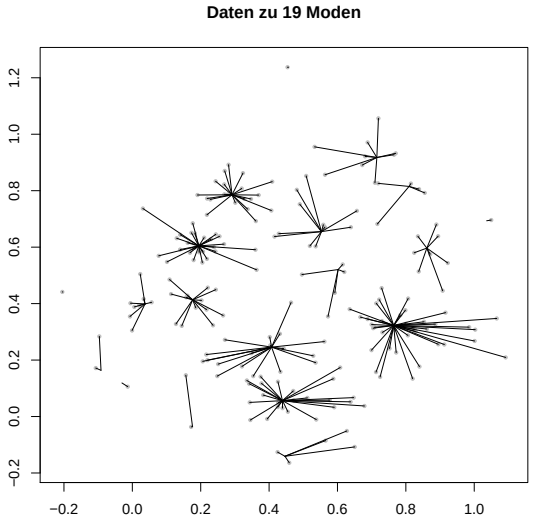

Abbildung 3.5: Zuordnung der Daten in Beispiel 3.1.1 zu 19 Modi im Pilot-Dichteschätzer beim Konstruieren von **V2**

in Abbildung 3.6, dass vier Komponenten des wahren Modells (rosa/schwarz/rot/grün eingefärbt) im **V2** identifiziert worden sind, während die gelbe Komponente in **V2** in zwei Teile gesplittet wurde. **V2** stellt einen besseren Dichteschätzer dar im Vergleich zu **V1**, weil es die Möglichkeit bietet, dass die Form einer Komponente im gemischten Modell in Beispiel 3.1.1 richtig geschätzt wird, und t_j den Mittelwert μ_j widerspiegelt. Die folgenden zwei Probleme von **V2** sind zu beachten:

1. **V2** hängt stark von dem Glättungsparameter des Pilot-Dichteschätzers ab;

2. Beim Konstruieren von **V2** werden die Daten zuerst zu den am nächsten liegenden Modi zugeordnet, was zu einer Partition der Daten führt, was impliziert, dass **V2** auch vom Partitionsverfahren abhängt.

Modell 3 Gemischtes Modell aus dem Model Based Clustering von Fraley & Raftery (2002)

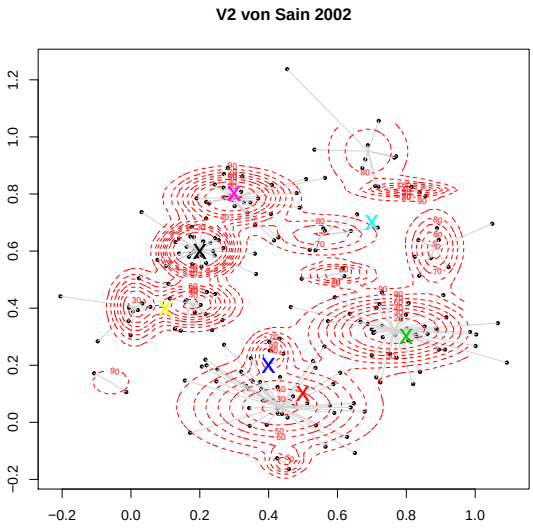

Abbildung 3.6: **V2** von Sain (2002) im Contour-Plot in Beispiel 3.1.1

Dabei wird die unbekannte wahre Dichte durch

$$\hat{f}_{fr}(x) = \sum_{j=1}^{m} w_j \phi(x, \theta_j), \tag{3.5}$$

geschätzt, wobei m die Anzahl der Komponenten im gemischten Modells ist, und w_j für die Wahrscheinlichkeit steht, dass x zu der j-ten Komponente gehört. Die Modellparameter w_j und θ_j werden durch den EM-Algorithmus bestimmt. Die Log-Likelihood Funktion sieht entsprechend wie folgt aus (Fraley & Raftery (2002)):

$$L_{fr}(\theta_1, ..., \theta_m; w_1, ..., w_m, z_{ij}|x) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} log(w_j \phi(X_i|\theta_j)) \tag{3.6}$$

wobei $z_{ij} = 1$, wenn X_i zu der j-ten Komponente gehört, und 0 ansonsten. Beim E-Step des EM-Algorithmuses schätzt man z_{ij} durch

$$\hat{z}_{ij} = \frac{w_j^{(t)} \phi(X_i|\theta_j^{(t)})}{\sum_{k=1}^{m} w_k^{(t)} \phi(X_i|\theta_k^{(t)})} \tag{3.7}$$

wobei $w_j^{(t)}$ und $\theta_j^{(t)}$ die geschätzten Modellparameter aus der letzten Iteration sind. Beim M-Step des EM-Algorithmuses wird die Log-Likelihood Funktion in (3.6) maximiert in Bezug auf w und θ unter Festhaltung von z_{ij}. Als Initialwert für das gemischte Modell

beim EM-Algorithmus wird das Resultat aus einem hierarchischen Clustering genommen. Eine ausführliche Beschreibung des Algorithmuses findet man in Celeux & Govaert (1995) und Fraley & Raftery (2002).

Beim Auswählen des für die Daten am besten geeigneten Modells wird das BIC-Kriterium benutzt mit

$$BIC_i = 2log(p(D|\hat{\theta}_i, M_i)) - v_i log(n), \tag{3.8}$$

wobei M_i das i-te Modell, D die vorgegebenen Daten, v_i die Anzahl der zu schätzenden unabhängigen Modellparameter in M_i. Die Hauptargumente für Nutzung des obigen BIC Kriteriums für die Auswahl des besten Modells sind (Fraley & Raftery (2002)):

- Man sucht das Modell mit maximaler $p(M_i|D)$ - die posterior Wahrscheinlichkeit von M_i gegeben Daten D;
- $p(M_i|D) \propto p(D|M_i)p(M_i)$ mit $p(D|M_i) = \int p(D|\theta_i, M_i)p(\theta_i|M_i)d\theta_i$, wobei $p(\theta_i|M_i)$ die a priori Verteilung von θ_i ist. Wenn wir $p(M_i)$, $i = 1, ..., K$ (K steht für die Anzahl der zu vergleichenden Modelle) als gleich annehmen können, dann ist das Modell mit maximaler $P(D|M_i)$ zu wählen;
- Es gilt $2log(p(D|M_i)) \approx BIC_i$ (Schwarz 1978; Haughton 1988; Fraley & Raftery 2002).

Abbildung 3.7 zeigt die BIC-Werte (definiert in (3.8)) von verschiedenen Modellen für die Daten in Beispiel 3.1.1. Man sieht in Abbildung 3.7, dass das beste Modell dem BIC Kriterium gemäß ein VII Modell ($\Sigma_j = \lambda_j I_2$) mit 6 Komponenten ist. In Abbildung 3.8 wird dieses beste Modell im Contour-Plot dargestellt. Man sieht in Abbildung 3.8, dass 6 Komponenten des wahren Modells (rosa/schwarz/gelb/rot/grün/leicht blau eingefärbt) dadurch identifiziert worden sind, während die blaue Komponente falsch zu der roten Komponente zugeordnet wurde. In unserem Beispiel stellt **Modell 3** den besten Schätzer (auch im Vergleich zu **Modell 4** unten) für die wahre Dichte dar - dies ist zu erwarten, weil die Daten aus einer gemischten Normalverteilung kommen. Falls diese Modellannahme nicht stimmt, kann das Model Based Clustering aber nicht immer ein zuverlässiges Resultat liefern.

Modell 4 Gemischte Modelle aus dem CEM- und SEM-Algorithmus

Das gemischte Modell $\hat{f}_n(x, \theta)$ konstruiert man wie folgt:

S1 Man berechnet einen Fest-Kerndichteschätzer mit Gaussian Kernfunktion $k(.)$ und Glättungsparameter h als Pilot-Kerndichteschätzer;

S2 Man führt den EM-Algorithmus auf das Modell aus S1 unter Festhaltung von w_i und Σ_i für $i = 1, ..., n$;

S3 Man ordnet die Daten zu den neuen Modi zu und verändert w_i, $i = 1, ..., n$ entsprechend;

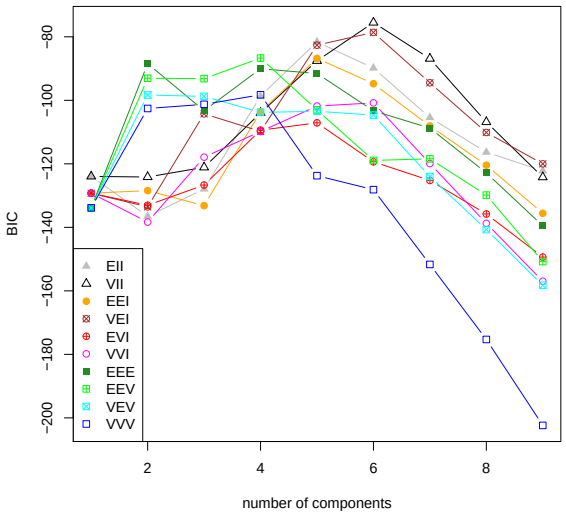

Abbildung 3.7: BIC-Werte der gemischten Modelle aus dem Model Based Clustering von Fraley & Raftery (2002) in Beispiel 3.1.1

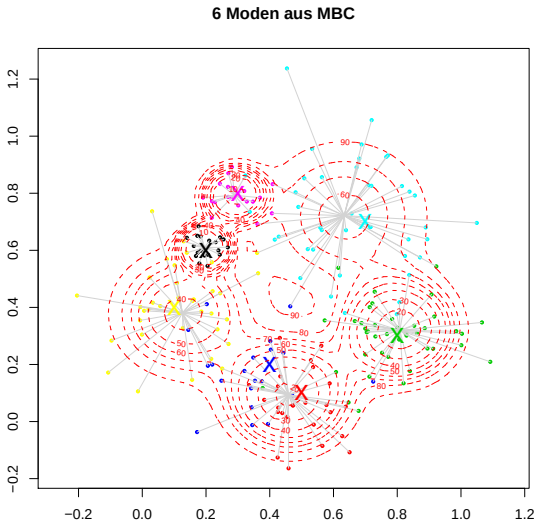

Abbildung 3.8: Das beste gemischte Modell (VII Modell mit 6 Komponenten) aus dem Model Based Clustering von Fraley & Raftery (2002) in Beispiel 3.1.1

S4 Man kombiniert die Einzel-Punkt-Cluster mit dem nächsten Cluster;

S5 Man gibt Σ_i einen Initialwert und führt den CEM- oder SEM-Algorithmus (Celeux & Govaert (1992)) durch.

Ein paar Erklärungen zu **Modell 4**:

1. Beim Bestimmen der Modellparameter für $\hat{f}_n(x, \theta)$ werden die EM (hier EM, CEM und SEM) Algorithmen benutzt. Im Vergleich zu den Modellen von Sain (2002) ist **Modell 4** sowohl für große Datensätze als auch für multivariate Daten besser geeignet, weil in solchen Fällen die Modellparameter im **V1** und **V2** nach dem LSCV-Kriterium beim numerischen Verfahren schwierig zu bestimmen sind;

2. **Modell 4** beruht auf der Grundidee des nichtparametrischen Clusterings, nämlich dass man die Daten nach ihrer Zuordnung zu den Modi der Dichtefunktion in Clustern aufteilen kann. Da die wahre Dichtefunktion $f(x, \theta)$ in der Regel unbekannt ist, wird $f(x, \theta)$ zuerst in S1 durch den Pilot Kerndichteschätzer mit Glättungsparameter h geschätzt;

3. Wenn man einen Kerndichteschätzer als ein gemischtes Modell betrachtet, dann besteht das Modell aus n Komponenten mit $w_i = 1/n$, $\mu_i = X_i$, $diag(\Sigma_i) = h^2$ für $i = 1, ..., n$, wobei $diag(A)$ für den Vektor der Diagonalelemente von Matrix A steht. Man führt einen EM-Algorithmus unter Festhaltung von w_i und Σ_i durch, um die n Komponenten nach der Modalstruktur des Pilot-Kerndichteschätzers zu fusionieren. In den EM-Algorithmus werden Zufallsvariablen $z_1, ..., z_n$ eingeführt mit $z_i = (z_{i1}, ..., z_{im})^T$, wobei $z_{ij} = 1$, wenn X_i von der j-ten Komponente erzeugt worden ist, and 0, sonst. Beim E-Step wird die posterior Wahrscheinlichkeit $P^t(j|X_i, \theta^{(t)})$, dass X_i von der j-ten Komponente erzeugt wird, folgendermaßen berechnet:

$$P^{(t)}(j|X_i, \theta^{(t)}) = \frac{w_j^{(t)} \phi(X_i|\theta_j^{(t)})}{\sum_{k=1}^{m} w_k^{(t)} \phi(X_i|\theta_k^{(t)})} \tag{3.9}$$

wobei $w_j = 1/n$, $m = n$ hier, und $\theta_j^{(t)}$ der Modellparameter aus der letzten Iteration des EM-Algorithmuses. Die Log-Likelihood Funktion mit $P^{(t)}(j|X_i, \theta^{(t)})$ sieht wie folgt aus:

$$LL(w, \theta; w^{(t)}, \theta^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{m} (logP(j) + log\phi(X_i|\theta_j)) P^{(t)}(j|X_i, \theta^{(t)}) \tag{3.10}$$

Beim M-Step wird der neue Mittelwert $\mu_j^{(t+1)}$ wie folgt bestimmt:

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} P^{(t)}(j|X_i, \theta^{(t)}) X_i}{\sum_{i=1}^{n} P^{(t)}(j|X_i, \theta^{(t)})} \tag{3.11}$$

und im Fall $m = 1$ gilt

$$\mu^{(t+1)} = \frac{\sum_{i=1}^{n} k((\mu^{(t)} - X_i)/h)X_i}{\sum_{i=1}^{n} k((\mu^{(t)} - X_i)/h)} \tag{3.12}$$

Eine ausführliche Beschreibung dieses EM-Algorithmuses findet man z.B. in McLachlan & Basford (1988), Celeux & Govaert (1995) und Nabney (2002). Die gleiche Formel wie in (3.12) wurde auch in Hinneburg et al. (2007) für die Bestimmung des nächsten Punkts x^{l+1} für x^l bei ihrem Hill-Climbing Algorithmus verwendet;

4. Mit dem obigen Verfahren landet man von X_i aus auf das am nächsten liegende lokale Maximum des Pilot Dichteschätzers. Die Clusterstruktur wird dadurch identifiziert, indem man X_i, $i = 1, ..., n$ zu dem entsprechenden Modus zuordnet;

5. Beim Konstruieren von $\hat{f}_n(x, \theta)$ spielt der Glättungsparameter h des Pilot-Kerndichteschätzers eine wichtige Rolle. Für die Bestimmung von h ist die LSCV-Methode zu empfehlen, weil sie eigentlich komplett daten-orientiert ist. Im hochdimensionalen Fall sind die Normal-Reference-Bandbreiten zu empfehlen aus den folgenden zwei Hauptgründen:

- Im hochdimensionalen Datenraum sind die Daten in der Regel dünn verteilt. Mit der LSCV-Methode erhält man in meisten Fällen einen relativ kleinen Glättungsparameter, so dass sich die Fusion bei S2 und S3 nicht lohnt;
- Außerdem sind die LSCV-Bandbreiten schwierig zu bestimmen, weil es zu viele Parameter beim numerischen Verfahren zu schätzen gibt.

6. In S4 werden die Einzel-Punkt-Cluster mit dem am nächsten liegenden Cluster kombiniert, weil die Existenz eines Einzel-Punkt-Clusters zur trivialen Lösung beim EM-Algorithmus führt (Nabney (2002));

7. Ein CEM (Classification EM) Algorithmus (Celeux & Govaert 1992) unterscheidet sich vom EM-Algorithmus dadurch, dass man einen C-Step zwischen dem E-Step und M-Step des EM-Algorithmuses hinzufügt. Beim C-Step verändert man die Partition der Daten anhand von $P^{(t)}(j|X_i, \theta^{(t)})$ aus dem E-Step so, dass man X_i zu der j-ten Komponente zuordnet mit maximaler $P^{(t)}(j|X_i, \theta^{(t)})$;

8. Ein SEM (Stochastic EM) Algorithmus (Celeux & Govaert 1992) unterscheidet sich vom EM-Algorithmus dadurch, dass man einen S-Step zwischen dem E-Step und M-Step des EM-Algorithmuses hinzufügt. Beim S-Step verändert man die Partition der Daten anhand von $P^{(t)}(j|X_i, \theta^{(t)})$ aus dem E-Step so, dass man X_i nach der posterior Wahrscheinlichkeit $P^{(t)}(j|X_i, \theta^{(t)})$ per Zufall zu der j-ten Komponente zuordnet;

9. Mit dem SEM-Algorithmus von Celeux & Govaert (1992) hat man die Möglichkeit, aus einem kleineren lokalen Maximum der Likelihood Funktion hinaus zu kommen und auf ein größeres lokales Maximum zu landen. Für diesen Zweck wird hier eine modifizierte Version dieses SEM-Algorithmuses benutzt, die man wie folgt formulieren kann:

- E-Step: Man berechnet die posterior Wahrscheinlichkeit $P^{(t)}(j|X_i, \theta^{(t)})$;
- SM-Step:

  SM1 Man ordnet X_i nach der posterior Wahrscheinlichkeit $P^{(t)}(j|X_i, \theta^{(t)})$ per Zufall zu der j-ten Komponente zu und schätzt $w_j^{(t+1)}$ und $\theta_i^{(t+1)}$ entsprechend;

  SM2 Man berechnet den neuen Wert der Log-Likelihood Funktion $LL(w^{(t+1)}, \theta^{(t+1)})$ und vergleicht ihn mit $LL(w^{(t)}, \theta^{(t)})$. Falls

  $$LL(w^{(t+1)}, \theta^{(t+1)}) < LL(w^{(t)}, \theta^{(t)})$$

  dann geht man zum E-Step zurück mit $w = w^{(t)}$ und $\theta = \theta^{(t)}$;

  SM3 Falls

  $$|LL(w^{(t+1)}, \theta^{(t+1)}) - LL(w^{(t)}, \theta^{(t)})|/LL(w^{(t+1)}, \theta^{(t+1)}) < eps$$

  stopp, falls nicht, dann zurück zum E-Step, wobei *eps* als Abbruchkriterium vorgegeben ist.

- V-Step: Man führt den CEM-Algorithmus durch auf Basis von dem Resultat aus dem SM-Step.

Mit dem obigen modifizierten SEM-Algorithmus landet man sicher auf ein großes lokales Maximum der Log-Likelihood Funktion. Er liefert in der Regel schon nach ein paar Versuchen ein besseres Resultat als das aus dem CEM-Algorithmus. Abbildung 3.9 zeigt 23 steigende Funktionswerte aus 500 Iterationen des SEM-Algorithmuses im Vergleich zu den Funktionswerten aus dem CEM-Algorithmus. Man sieht in Abbildung 3.9, dass man mit dem modifizierten SEM-Algorithmus auf ein größeres lokales Maximum landet.

10. $\hat{f}_n(x, \theta)$ bietet nur eine Version der Daten, weil das Modell stark von dem Pilot-Kerndichteschätzer abhängt.

In Abbildung 3.10 werden die gemischten Modelle aus dem CEM-Algorithmus (Grafik links) und modifizierten SEM-Algorithmus (Grafik rechtes) in Contour-Linien dargestellt. Man sieht in Abbildung 3.10, dass der CEM-Algorithmus ein ähnliches Resultat liefert wie **V2** von Sain (2002), während man mit dem modifizierten SEM-Algorithmus ein besseres Ergebnis bekommt. Das gemischte Modell aus dem modifizierten SEM-Algorithmus zeigte 5 Komponenten richtig (rosa/schwarz/gelb/rot/grün eingefärbt), während die blaue Komponente falsch zu der roten Komponente zugeordnet worden ist und die leicht-blaue Komponente in 5 kleinen Subkomponenten gesplittet wurde.

In Tabelle 3.1 stellt man die entsprechenden Likelihood, BIC und ARI (Adjusted Rand Index) der obigen Modelle dar. Der Adjusted Rand Index (Hubert & Arabie (1961)) misst die Ähnlichkeit vom obigen Resultat aus dem Dichteschätzer basierten Clustering mit der vorgegebenen Klassifikation der Daten und wurde wie folgt definiert: Seien P_1

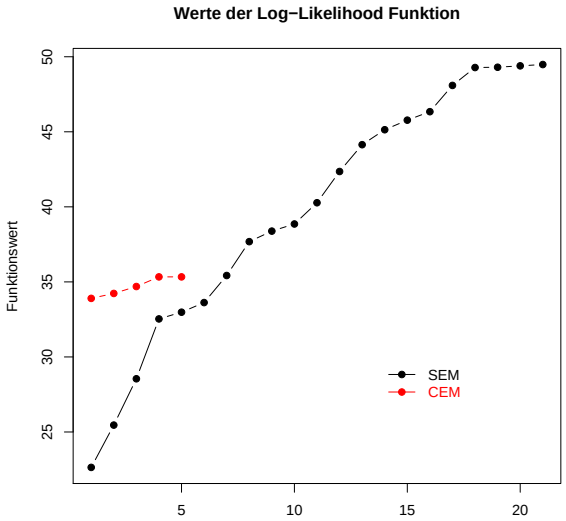

Abbildung 3.9: 23 steigende Funktionswerte der Log-Likelihood aus 500 Iterationen des modifizierten SEM-Algorithmuses in schwarzen Punkten und Funktionswerte der Log-Likelihood aus dem CEM-Algorithmus in roten Punkten in Beispiel 3.1.1

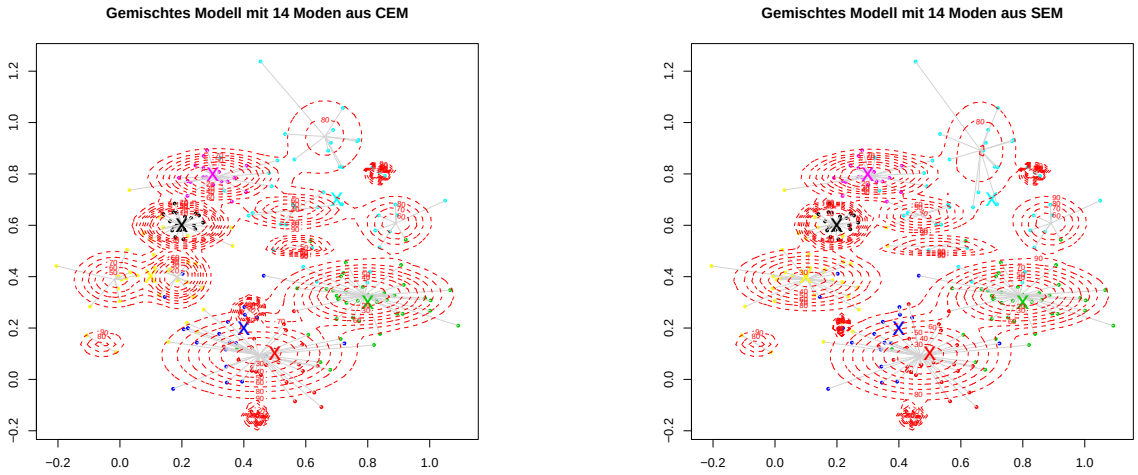

Abbildung 3.10: Gemischtes Modell aus dem CEM-Algorithmus in Contour-Linien (Grafik links) und Gemischtes Modell aus dem modifizierten SEM-Algorithmus in Contour-Linien (Grafik rechts) in Beispiel 3.1.1

	KDE	**V1**	**V2**	MBC	CEM	SEM
Likelihood	56,8002	-19,7611	18,2840	23,1776	-51,5289	57,4882
BIC	-3070,688	-749,4967	-329,0158	-75,5061	-468,6417	-250,6074
ARI	0	0,0981	0,4421	0,6498	0,4303	0,4767

Tabelle 3.1: Vergleich der 6 Varianten des Modells in (3.2) bezüglich Likelihood, BIC und ARI anhand des Beispiels 3.1.1

und P_2 zwei Partitionierungen von n Objekten, und n_{ij} die Anzahl der Objekte, die zu der i-ten Submenge in P_1 und gleichzeitig zu der j-ten Submenge in P_2 gehören, dann

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2} / \binom{n}{2}}{\frac{1}{2}\left(\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right) - \sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2} / \binom{n}{2}}, \tag{3.13}$$

wobei $n_{i.} = \sum_j n_{ij}$ und $n_{.j} = \sum_i n_{ij}$.

Aus Vergleich der obigen Modelle durch die grafische bzw. tabellarische Darstellung sieht man, dass das Modell aus dem Model Based Clustering die wahre Dichte in Beispiel 3.1.1 am besten widerspiegelt. In der Regel liefert ein parametrisches Modell ein sinnvolles Resultat, wenn die Modellannahme getroffen ist. Zum Schluss dieses Abschnitts noch ein paar Bemerkungen zum EM-Algorithmus und gemischten Modell:

- Das Resultat des EM-Algorithmuses hängt stark vom Startwert ab. Als Startwert für den EM-Algorithmus nimmt man beispielsweise beim Model Based Clustering (**Modell 3**) das Resultat aus einem hierarchischen Clustering. Es besteht durchaus die Möglichkeit, dass man das Resultat aus anderen Clusteringverfahren dafür verwendet;

- Beim EM-Algorithmus besitzt die Zielfunktion in der Regel mehrere lokale Maxima. Es ist immer fraglich, wie man ein geeignetes lokales Maximum wählt. Mit dem (modifizierten) SEM-Algorithmus hat man z.B. die Möglichkeit, mehrere lokale Maxima durch die stochastische Zuordnung beim (SM) S-Step zu finden und das größte Maximum aus ihnen auszuwählen;

- Beim EM-Algorithmus besteht die Möglichkeit, dass man auf das triviale Maximum ($LL(.)$ gegen unendlich) landet, wenn es Einzel-Punkt-Cluster gibt. Um das Problem zu beheben, wird in diesem Abschnitt der Einzel-Punkt-Cluster vor dem M-Step mit dem am nächsten liegenden Cluster fusioniert, wenn er durch den C- bzw. (SM) S-Step beim CEM- bzw. (modifizierten)) SEM-Algorithmus erzeugt worden ist. Für dieses Problem gibt es verschiedene Lösungsvorschläge. In Nabney (2002) wird z.B. das Problem dadurch gelöst, indem der Varianz Σ des Einzel-Punkt-Clusters beim M-Step die gesamte Varianz der Daten zugewiesen wird;

- Es gibt in der Regel keine 1 zu 1 Relation zwischen den Modi und Modellkomponenten. Es ist üblich, dass ein Modus aus mehreren Modellkomponenten zusammengebastelt wird;
- Dass ein parametrisches Verfahren ein sinnvolles Resultat liefert, wenn die Modellannahme getroffen wird, ist äquivalent dazu, dass in diesem Fall man mit einem nichtparametrischen Verfahren ein relativ schlechteres Ergebnis bekommt, so wie es hier in diesem Abschnitt zu sehen ist.

3.2 Dichteschätzer basierte Clusteranalyse

Die in diesem Abschnitt vorzustellende Clusteranalyse beruht auf den folgenden zwei Annahmen:

- Die empirischen Daten sind aus einer Wahrscheinlichkeitsverteilung mit Dichtefunktion f;
- Die Daten können zu den Domänen der Modi in f zugeordnet werden.

Der Zusammenhang von der Multimodalität und Datengruppierung wurde zuerst in der Arbeit „Mode Analysis“ von Wishart (1969) untersucht. Hartigan (1975) erweiterte die Idee von Wishart und führte den Begriff „High Density Clusters“ in die Dichteschätzer basierte Clusteranalyse ein, die als zusammenhängende Komponenten der λ-Niveaumenge beschrieben wurden. Eine λ-Niveaumenge wurde in dieser Arbeit als $L(\lambda, f)$ geschrieben und wie folgt definiert:

$$L(\lambda, f) = \{x | f(x) > \lambda\} \tag{3.14}$$

Da f in der Praxis meist unbekannt ist, schätzt man f in der Regel durch einen geeigneten Dichteschätzer $\hat{f}$ und zieht die Modalstruktur von $\hat{f}$ in Betracht für das Clustering der empirischen Daten. Die Grundidee von Hartigan liegt darin, dass falls die Daten eine Clusterstruktur haben, dann sollte diese Struktur in den Hohe-Dichte-Regionen auffällig und gut zu identifizieren sein. Ausgehend von diesem Grundgedanken wurden in den letzten Jahren verschiedene statistischen Verfahren entwickelt, die auf der Modalanalyse der Single- und Multi-Niveaumenge beruhen. Bei Methoden auf Basis der Single-Niveaumenge wird die Clusterstruktur der Daten in $L(\lambda, f)$ mit nur einem λ untersucht, typische Methoden befinden sich z.B. in Ester et al. (1996), Walther (1997) und Cuevas et al. (2000, 2001). Der Hauptnachteil dieser Methode liegt in den folgenden zwei Punkten:

- Man untersucht nicht die ganze Clusterstruktur der Daten;
- Das Resultat wird stark von der Auswahl von λ beeinflusst.

Um dieses Problem zu beheben, wurden Methoden auf Basis der Multi-Niveaumenge vorgeschlagen, wobei die Clusterstruktur der Daten über verschiedenen Dichteniveaus geschätzt wird, typische Methoden sind z.B. Runt Pruning von Stuetzle (2003), Level Set Tree von Klemelä (2004, 2005) und die Generalized Single Linkage Methode von Stuetzle et al. (2007). Zwei Punkte sind hier zu bemerken:

- Das Resultat aus allen beiden (Single- und Multi-Niveaumenge basierten) Verfahren hängt eigentlich von Dichteschätzer $\hat{f}$ ab;
- Es ist in der Regel unvermeidbar, dass manche Modi in $\hat{f}$ Artefakte wegen der Datenerhebung sind.

In diesem Abschnitt werden zuerst die DBSCAN Methode, die Methoden von Stuetzle (2003) und Stuetzle et al. (2007) anhand des Beispiels 3.1.1 vom letzten Abschnitt vorgestellt und grafisch veranschaulicht. Dann stellt man die Methode von Duong et al. (2007) vor, mit der man signifikante Modi identifizieren kann. Im Folgenden wird immer angenommen, dass $X_1, ..., X_n \in R^d$ eine unabhängige identisch verteilte Stichprobe aus einer Wahrscheinlichkeitsverteilung mit Dichtefunktion f sind.

DBSCAN von Ester et al. (1996)

In die DBSCAN-Clusteranalyse wurden die folgenden Begriffe eingeführt:

- $N_\epsilon(p)$ (ϵ-Nachbarschaft von P): Seien D die Daten aus R^d, p und q zwei beliebige Punkte aus D, dann
$$N_\epsilon(p) = \{q \in D | dist(p, q) \leq \epsilon\} \tag{3.15}$$
wobei $dist(p, q)$ ein Distanzmaß von p und q ist;
- $MinPts$: Minimale Anzahl von Punkten in $N_\epsilon(p)$;
- Kernpunkte (core points): Punkte innerhalb eines Clusters;
- Randpunkte (border points): Punkte am Rand eines Clusters;
- Direkt Dichte-Erreichbar (directly density-reachable): p ist von q aus direkt dichte-erreichbar, wenn $p \in N_\epsilon(q)$ und $|N_\epsilon(q)| \geq MinPts$, wobei $|N_\epsilon(q)|$ für die Anzahl der Daten in $N_\epsilon(q)$ steht;
- Dichte-Erreichbar (density-reachable): p ist von q aus dichte-erreichbar, wenn es eine Reihe von Daten $p_1, ..., p_k$ mit $p_1 = q$ und $p_k = p$ gibt, dass p_{i+1} von p_i aus direkt dichte-erreichbar ist;
- Dichte-Zusammenhängend (density-connected): p und q sind dichte-zusammenhängend, wenn es einen Punkt o existiert, dass beide Punkte p und q von o aus dichte-erreichbar sind.

Aufgrund der obigen Definitionen wird ein Cluster C definiert als eine nichtleere Submenge von D mit folgenden zwei Eigenschaften:

1. $\forall p, q \in D$: wenn $p \in C$ und q von p aus dichte-erreichbar ist, dann $q \in C$;
2. $\forall p, q \in C$: p und q sind dichte-zusammenhängend.

Wenn *dist* z.B. die euklidische Distanz bedeutet, dann kann man mit der DBSCAN Methode die „High Density Cluster" identifizieren, die aus allen X_i mit $\hat{f}_{uni}(X_i, h) \geq \lambda$, wobei $\hat{f}_{uni}(X_i, h)$ für Kerndichteschätzer mit der sphärischen Uniform-Kernfunktion und Glättungsparameter $h = \epsilon I_d$ steht und $\lambda = MinPts/(n(2\epsilon)^d)$, und denjenigen Punkten, deren Distanz zu X_i kleiner ϵ sind, bestehen.

Man sieht in der obigen Beschreibung der DBSCAN Methode, dass deren Resultat von ϵ und $MinPts$ abhängt. In unserem Beispiel setzt man $MinPts = 4$ (wie in Ester et al. (1996)), $\epsilon \in \{0,075; 0,080; 0,085; 0,090\}$ und stellt die Resultate in Abbildung 3.11 dar, wobei die Störungsbeobachtungen in grau gezeichnet werden und die Daten aus verschiedenen High Density Clustern unterschiedlich eingefärbt sind. Man sieht in Abbildung 3.11, dass die DBSCAN Methode in unserem Beispiel eigentlich schlechtere Resultate liefert im Vergleich zu den Methoden in Abschnitt 3.1, z.B., in der Grafik oben links ($\epsilon = 0,075$) hat man zwar 29% der Daten als Störungsbeobachtungen klassifiziert, kann aber die zwei Komponenten (oben links in der Grafik) des wahren gemischten Modells (vgl. Abs. 3.1) nicht trennen; in der Grafik unten rechts ($\epsilon = 0,090$) hat man wenigere Störungsbeobachtungen (11% der Daten), aber dafür einen großen in schwarz eingefärbten Cluster, der 78,5% der Daten enthält.

Runt Pruning von Stuetzle (2003)

Die Grundidee der Runt Pruning Methode besteht in der Schätzung vom Cluster Baum von f (hier als CB geschrieben). CB bezieht sich auf die hierarchische Struktur von f und wird wie folgt definiert:

1. Jeder Knoten N von CB steht für eine Submenge $T(N) = \{x | f(x) > \lambda(N)\}$ von $T = \{x | f(x) > 0\}$. Der Wurzel von CB steht für T;
2. Beim Bestimmen der Subknoten von N findet man das kleinste λ_s, so dass die Menge $L(\lambda_s; f) \cap T(N)$ k paarweise disjunkte zusammenhängende Komponenten $S_1, ..., S_k$ enthält. Dabei gibt es 3 Fälle zu unterscheiden:
   a) Falls λ_s nicht existiert, dann ist N ein Blatt von CB;
   b) Falls $k = 2$, dann hat N zwei Subknoten N_1 und N_2, die für S_1 und S_2 stehen;
   c) Falls $k > 2$, dann hat N zwei Subknoten N_1 und N_2, die für $S_1' = S_1$ und $S_2' = S_2 \cup ... \cup S_k$ stehen;
3. Jedes Blatt von CB steht für einen Cluster.

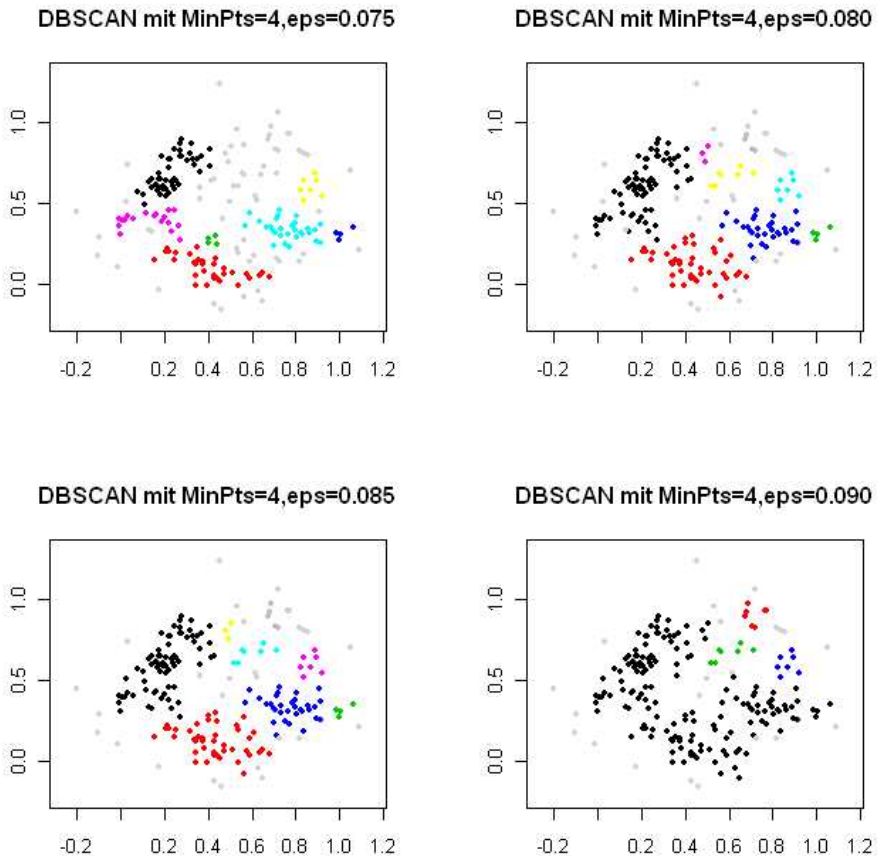

Abbildung 3.11: Clustering der Daten in Beispiel 3.1.1 mit der DBSCAN Methode mit $MinDts = 4$ und $\epsilon \in \{0,075; 0,080; 0,085; 0,090\}$. Daten aus verschiedenen High Density Clustern unterschiedlichen eingefärbt. Störungsbeobachtungen in grau gezeichnet.

Da f in der Regel unbekannt ist, wurde CB in Stuetzle (2003) durch den Cluster Baum von $\hat{f}_{nn}$ (hier als $\widehat{CB}_{nn}$ geschrieben.) geschätzt. $\hat{f}_{nn}$ ist der Nearest Neighbour Dichteschätzer mit

$$\hat{f}_{nn}(y) = \frac{1}{nVd(y,x)^d} \tag{3.16}$$

wobei V für das Volumen der Einheitskugel in R^d steht und $d(y,x) = min_i d(y, X_i)$, $i \in \{1, ..., n\}$. $\hat{f}_{nn}(y)$ und $\widehat{CB}_{nn}$ besitzen die folgenden Eigenschaften:

- $\hat{f}_{nn}(y)$ hat eine große Varianz;
- $\hat{f}_{nn}(y)$ hat Singularitäten an $X_1, ..., X_n$;
- $$L(\lambda; \hat{f}_{nn}) = \{x | \hat{f}_{nn}(x) > \lambda\} = \bigcup_i K_u(X_i, r(\lambda)) \tag{3.17}$$

  wobei K_u für die Kugel mit Zentrum X_i und Radius $r(\lambda)$ steht und $r(\lambda) = (\frac{1}{nV\lambda})^{\frac{1}{d}}$ (aus (3.16));
- Sei $G = (V, E)$ ein vollständiger Graph mit $V = D$, $E = \{e_{ij}\}$ mit $i, j = 1, ..., n$ und $i \neq j$, $w(e_{ij}) = w(X_i, X_j) = ||X_i - X_j||_2$, wobei $w(e_{ij})$ für das Gewicht der Kante e_{ij} in E steht, und weiter seien B der Minimale Erzeugende Baum von G und $B_1, ..., B_k$ paarweise disjunkte Subbäume von B, deren allen Kanten ihr Gewicht größer $r(\lambda)$ haben, und seien $D_1, ..., D_k$ die entsprechenden Submengen von D in $B_1, ..., B_k$, dann sind $L_i = \bigcup_{X_j \in D_i} K_u(X_j, r(\lambda))$, $i = 1, ..., k$ die paarweise disjunkten zusammenhängenden Komponenten von $L(\lambda; \hat{f}_{nn})$;
- $\widehat{CB}_{nn}$ hat n Blätter;
- $\widehat{CB}_{nn}$ ist isomorph zum Single Linkage Dendrogramm. Die linke Grafik in Abbildung 3.12 zeigt das Single Linkage Dendrorgramm für die Daten in Beispiel 3.1.1;
- $\widehat{CB}_{nn}$ kann durch rekursive Abbrüche der Kanten (ihrem Gewicht nach) in B erhalten werden. Die rechte Grafik in Abbildung 3.12 zeigt den Minimalen Erzeugenden Baum für die Daten in Beispiel 3.1.1.

Da ein $\widehat{CB}_{nn}$ n Blätter hat und deswegen die Clusterstruktur der Daten nicht widerspiegelt, schneidet man $\widehat{CB}_{nn}$ dann nach dem „Runt Size“, damit man signifikante Cluster identifizieren kann. Dieses Pruning-Verfahren kann wie folgt grob beschrieben werden:

- Der Runt Size eines Knotens N in $\widehat{CB}_{nn}$ wird definiert als die minimale Anzahl der Blätter in seinen beiden Subknoten. Da ein Nicht-Blatt-Knoten N in $\widehat{CB}_{nn}$ eine Kante e_{ij} in B bedeutet, ist der Runt Size von N in $\widehat{CB}_{nn}$ gleich der minimalen Anzahl der Daten in den beiden durch e_{ij} verbundenen Komponenten in B;

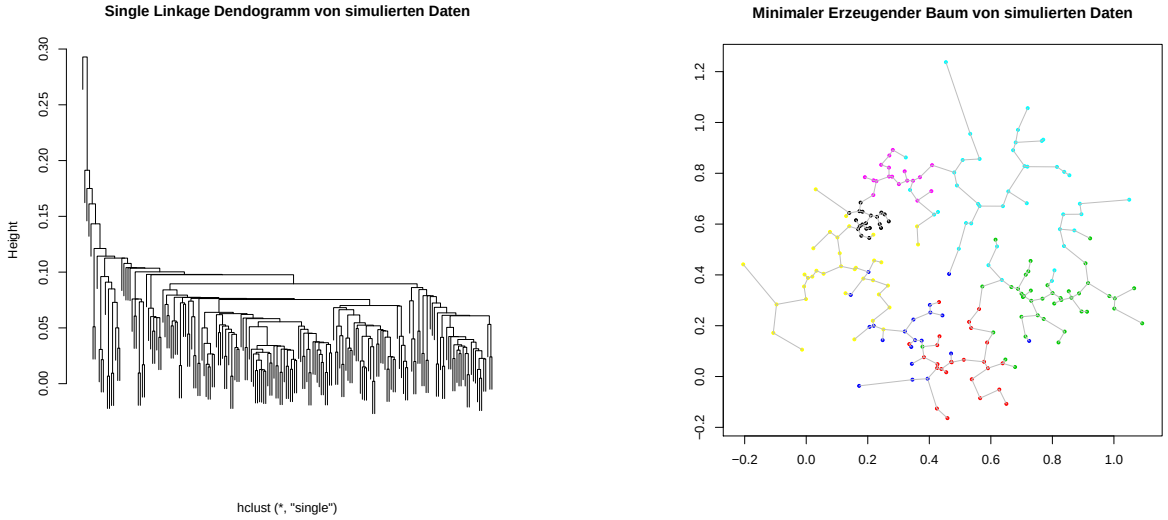

Abbildung 3.12: Single Linkage Dendrogramm (Grafik links) und Minimaler Erzeugender Baum (Grafik rechts) für die Daten in Beispiel 3.1.1

- Man berechnet den Runt Size für jeden Knoten und bestimmt eine Schwelle ϵ nach dem Quasi-Ellenbogen-Kriterium, das wie folgt aussieht: Abbildung 3.13 zeigt die 20 größten Runt Sizes der Knoten in $\widehat{CB}_{nn}$ für die Daten in Beispiel 3.1.1. Nach dem Quasi-Ellenbogen-Kriterium wird $\epsilon = 37$ ausgewählt, weil der Punkt 37 die größte „Gap“ zu den unteren Punkten hat;
- Diejenige Knoten, deren Runt Size größer als ϵ ist, werden gesplittet;
- Jedes Blatt im bearbeiteten $\widehat{CB}_{nn}$ steht für einen signifikanten „High Density Cluster“ in den Daten;
- Man ordnet die Daten dem Minimalen Erzeugenden Baum entsprechend zu den signifikanten „High Density Clustern“ zu, um die Clusterstruktur der ganzen Daten zu bekommen.

Mit $\epsilon = 37$ werden die Daten in Beispiel 3.1.1 in drei Clustern aufgeteilt. In der linken Grafik in Abbildung 3.14 werden die Daten aus den 3 Clustern in 3 verschiedenen Formen („1“ , „2“ und „3“) im Scatterplot gezeichnet. Die rechte Grafik in Abbildung 3.14 zeigt den entsprechenden geschätzten Cluster Baum aus Stuetzle (2003). Um die Clusterstruktur aus dem Runt Pruning Verfahren von der wahren Clusterstruktur der Daten zu unterscheiden, werden im Scatterplot in Abbildung 3.14 die Daten aus den wahren Clustern unterschiedlich eingefärbt. Man sieht im Scatterplot in Abbildung 3.14, dass das Clustering aus dem Runt Pruning Verfahren mit $\epsilon = 37$ die wahre Clusterstruktur nicht widerspiegelt, weil nur die in grün eingefärbte Komponente fast richtig identifiziert wurde. In Abbildung 3.13 sieht man, dass es eine andere Möglichkeit besteht, $\epsilon = 11$ für das

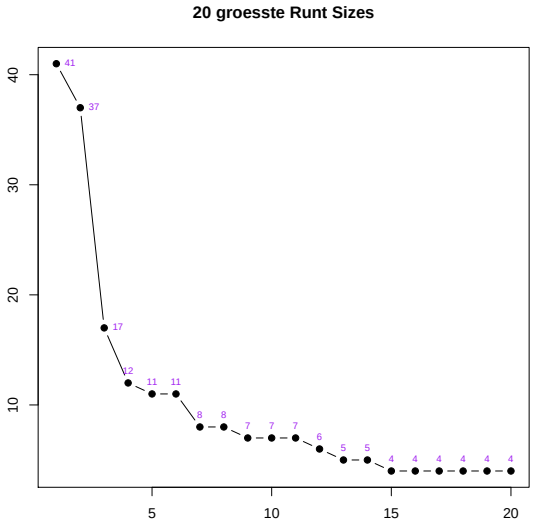

Abbildung 3.13: 20 größte Runt Sizes der Knoten in $\widehat{CB}_{nn}$ für die Daten in Beispiel 3.1.1

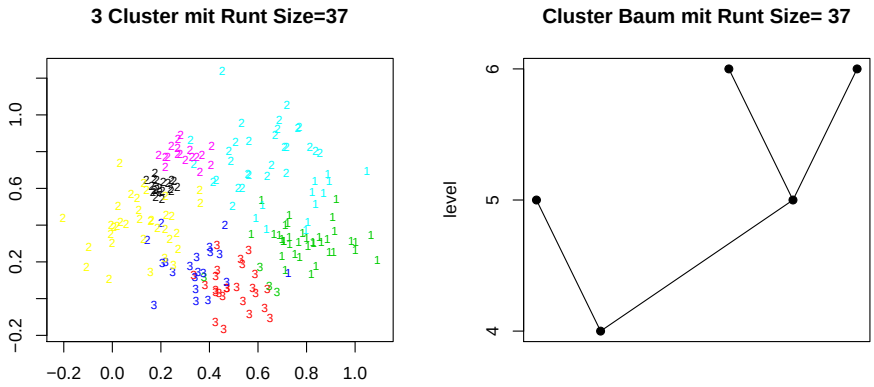

Abbildung 3.14: Daten in Beispiel 3.1.1 aus 3 verschiedenen Clustern aus dem Runt Pruning Verfahren in 3 unterschiedlichen Formen („1“ , „2“ und „3“) im Scatterplot (Grafik links) und der entsprechende geschätzte Cluster Baum aus Stuetzle (2003) (Grafik rechts)

Pruning Verfahren auszuwählen. Mit $\epsilon = 11$ werden die Daten in 7 Clustern aufgeteilt. Analog wie in Abbildung 3.14 wird diese Clusterstruktur in Abbildung 3.15 dargestellt. Man sieht in Abbildung 3.15, dass dieses Clustering die wahre Clusterstruktur besser

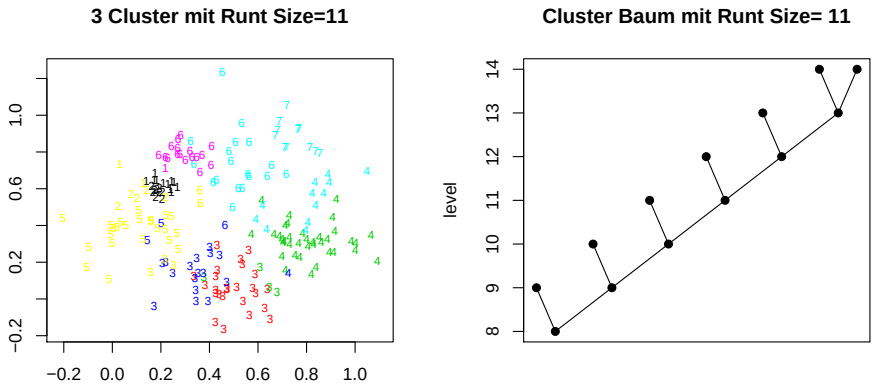

Abbildung 3.15: Daten in Beispiel 3.1.1 aus 7 verschiedenen Clustern aus dem Runt Pruning Verfahren in 7 unterschiedlichen Formen im Scatterplot (Grafik links) und der entsprechende geschätzte Cluster Baum aus Stuetzle (2003) (Grafik rechts)

schätzt als das Clustering mit $\epsilon = 37$. Da der Unterschied zwischen der geschätzten Clusterstruktur und wahren Clusterstruktur direkt in der Grafik zu erkennen ist, geht man darauf nicht weiter ein.

Generalized Single Linkage Methode von Stuetzle et al. (2007)

In Stuetzle et al. (2007) wurde $\widehat{CB}$ durch einen Graph Cluster Baum (hier als GCB geschrieben) approximiert, der wie folgt konstruiert wurde:

1. Man berechnet einen Dichteschätzer $\hat{f}$;
2. Man konstruiert einen vollständigen Graph $G = (V, E)$ mit $V = D$, $E = \{e_{ij}\}$ mit $i, j = 1, ..., n$ und $i \neq j$, $w(e_{ij}) = w(X_i, X_j) = min_{t\in[0,1]}\hat{f}((1-t)X_i + tX_j)$, wobei $w(e_{ij})$ für das Gewicht der Kante e_{ij} in E steht;
3. Ein Begriff „Threshold Graph“ $G(\lambda)$ wurde eingeführt, der als Subgraph von G aus denjenigen Kanten von G besteht, deren Gewicht größer λ ist, definiert wurde;
4. Ein GCB bezieht sich auf die Baumstruktur der zusammenhängenden Komponenten von $G(\lambda)$ mit verschiedenen λ;
5. Jeder Knoten N von GCB steht für einen Subgraph $T(N)$ von G, der mit einem Dichteniveau $\lambda(N)$ verbunden ist;

6. Der Wurzel von GCB steht für G mit $\lambda(G) = 0$;
7. Beim Bestimmen der Subknoten von N findet man das kleinste λ_s, so dass $G(\lambda_s) \cap T(N)$ k paarweise disjunkte zusammenhängende Komponenten $S_1, ..., S_k$ enthält. Dabei gibt es 3 Fälle zu unterscheiden:
   a) Falls λ_s nicht existiert, dann ist N ein Blatt von GCB;
   b) Falls $k = 2$, dann hat N zwei Subknoten N_1 und N_2, die für S_1 und S_2 stehen;
   c) Falls $k > 2$, dann hat N zwei Subknoten N_1 und N_2, die für $S_1' = S_1$ und $S_2' = S_2 \cup ... \cup S_k$ stehen;
8. Jedes Blatt von GCB steht für einen Cluster.

Ein GCB besitzt die folgenden Eigenschaften:

- Wenn X_i, X_j (i, j beliebig, $i \neq j$) in der selben zusammenhängenden Komponente in $G(\lambda)$ liegen, dann liegen X_i, X_j auch in der selben zusammenhängenden Komponente in $L(\lambda; \hat{f})$;
- X_i, X_j gehören zu der selben Komponente in $G(\lambda)$, dann und genau dann wenn sie zu der selben Komponente in $B_{max}(\lambda)$ gehören, wobei $B_{max}(\lambda)$ für den Subraum des Maximalen Erzeugenden Baums B_{max} steht, deren allen Kanten ihr Gewicht größer λ haben. Abbildung 3.16 zeigt den Maximalen Erzeugenden Baum für die Daten in Beispiel 3.1.1, wobei der Kerndichteschätzer mit LSCV-Bandbreiten als $\hat{f}$ verwendet wird.

Da ein GCB n Blätter hat und deswegen die wahre Clusterstruktur der Daten nicht widerspiegelt, schneidet man GCB analog wie oben beim Schneiden von $\widehat{CB}_{nn}$ beim Runt Pruning Verfahren aber nach dem Runt Excess Maß anstatt des Runt Sizes. Das Runt Excess Maß eines Knotens N im GCB wird definiert als das minimale Excess Maß seiner beiden Subknoten. Das Excess Maß eines Knotens N_s im GCB wird geschätzt durch

$$\hat{E}(N_s) = \frac{1}{n} \sum_i I(X_i \in T(N_s))(1 - \lambda(N_s)/\hat{f}(X_i)) \tag{3.18}$$

wobei I eine Indikatorfunktion ist mit $I(X_i \in T(N_s)) = 1$, falls $X_i \in T(N_s)$, und 0 ansonsten. Man verwendet das Excess Maß eines Knotens beim Pruning von GCB, um die signifikanten Cluster identifizieren zu können, weil sie in der Regel ein großes Excess Maß haben. Zu bemerken ist, dass der Übersichtlichkeit halber das Excess Maß im späteren Teil dieser Arbeit als

$$\hat{E}(N_s) = \sum_i I(X_i \in T(N_s))(1 - \lambda(N_s)/\hat{f}(X_i)) \tag{3.19}$$

gezeigt wird, weil ansonsten die Zahlen viel zu klein sind. In dieser Arbeit benutzt man dann ϵ für $\hat{E}(N_s)$ beim Pruning des Dendrogramms.

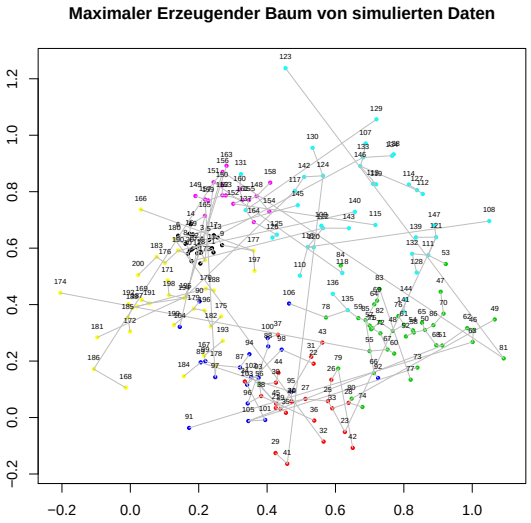

Abbildung 3.16: Maximaler Erzeugender Baum auf Basis von $\hat{f}$ für die Daten in Beispiel 3.1.1

Abbildung 3.17 zeigt die 20 größten Runt Excess Maß der Knoten (die letzten 5 Nullen sind aus der Rundung) im GCB für die Daten in Beispiel 3.1.1. Wenn man $\epsilon = 14,4$ nach dem Quasi-Ellenbogen-Kriterium auswählt und den GCB dementsprechend schneidet, dann werden die Daten in drei Clustern aufgeteilt. Abbildung 3.18 zeigt den geschätzten Cluster Baum, wobei die Ziffer neben dem Knoten für die Anzahl der Daten in der diesem Knoten entsprechenden Submenge der Daten steht. In Abbildung 3.19 werden die Daten aus den 3 Clustern im Scatterplot dargestellt. Man sieht in Abbildung 3.19, dass das Resultat aus der Generalized Single Linkage Methode mit $\epsilon = 14,4$ ähnlich wie das beim obigen Runt Pruning Verfahren mit $\epsilon = 37$ ist und die wahre Clusterstruktur nicht gut widerspiegelt. In Abschnitt 3.3 geht man auf die Generalized Single Linkage Methode von Stuetzle et al. (2007) weiter ein.

Methode von Duong et al. (2007)

Die Methoden von Chaudhuri & Marron (1999), Godtliebsen et al. (2002) und Duong et al. (2007) beruhen auf der Grundidee der Scale Space Theorie (Lindeberg 1994) aus der Computer Vision, wobei es eigentlich um die Repräsentation eines Bildes durch die Darstellung einer Serie von geglätteten Bildern geht. Die Anwendung der Grundidee der Scale Space Theorie im Bereich der nichtparametrischen Dichteschätzung liegt darin, dass man f durch Dichteschätzung mit einer Reihe von Glättungsparametern schätzt und die Eigenschaften von f anhand dieser Dichteschätzer untersucht, während der Schwerpunkt der meisten statistischen Methoden in diesem Bereich an der Suche nach einem „optimalen“ Schätzer von f liegt. Diese Vorgehensweise der Dichteschätzung ist besonders

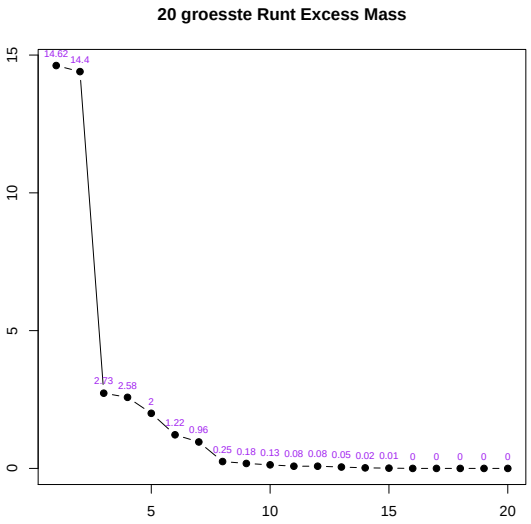

Abbildung 3.17: 20 größte Runt Excess Maße der Knoten im GCB für die Daten in Beispiel 3.1.1

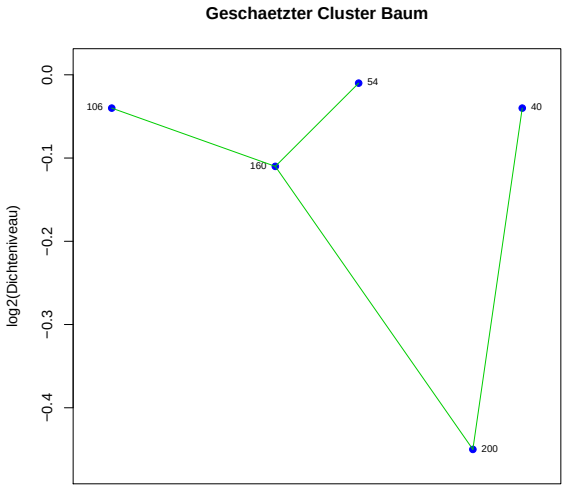

Abbildung 3.18: Geschätzter Cluster Baum aus der Generalized Single Linkage Methode von Stuetzle et al. (2007) für die Daten in Beispiel 3.1.1

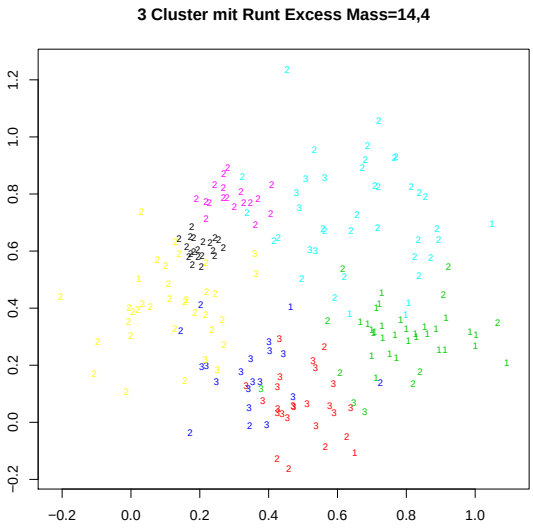

Abbildung 3.19: Daten in Beispiel 3.1.1 aus 3 Clustern aus der Generalized Single Linkage Methode von Stuetzle et al. (2007) in „1“, „2“ und „3“im Scatterplot

nützlich und hilfreich in der explorativen Datenanalyse, weil man davon ausgehen kann, dass „different levels of smoothing may reveal different useful information“ (Marron & Chung (1997)). Das zentrale Problem bei diesem Verfahren ist: Welche Eigenschaften sind zu finden und welche gefundenen Eigenschaften sind tatsächlich vorhanden.

In multivariater Datenanalyse bezieht sich die Eigenschaft der Daten in erster Linie auf das lokale Maximum in f, das man durch Berechnen von $\hat{f}'$ und $\hat{f}''$ schätzen kann. Die Verteilungseigenschaften von $\hat{f}'$ und $\hat{f}''$ wurden in Chaudhuri & Marron (1999), Godtliebsen et al. (2002) und Duong et al. (2007) genutzt, um zu testen, ob die gefundenen lokalen Maxima signifikant sind. Im Folgenden wird die Methode von Duong et al. (2007) anhand des Beispiels 3.1.1 vorgestellt.

Sei $H = (h_{ij})_{i,j\in\{1,\dots,d\}} \in R^{d\times d}$ die Bandbreitenmatrix mit $h_{ij} = h_i h_j$, dann gilt:

$$\hat{f}^{(r)}(x;H) = n^{-1}\sum_{i=1}^{n} K_H^{(r)}(x - X_i) \tag{3.20}$$

wobei $f^{(r)}$ für die r-te Ableitung von f steht, und

$$K_H'(x) = |H|^{-1/2}H^{-1/2}K'(H^{-1/2}x) \tag{3.21}$$

$$K_H''(x) = |H|^{-1/2}H^{-1/2}K''(H^{-1/2}x)H^{-1/2} \tag{3.22}$$

Unter gewissen Bedingungen (Duong et al. (2007)) haben $\hat{f}'(x;H)$ und $\hat{f}''(x;H)$ die folgenden asymptotischen Verteilungseigenschaften, wenn $n \to \infty$:

- $$n^{1/2}|H|^{1/4}H^{1/2}(\hat{f}'(x;H) - f'(x)) \xrightarrow{D} N(0, \Sigma'(x)) \tag{3.23}$$
  wobei $\xrightarrow{D}$ die Konvergenz in Verteilung bedeutet und $\Sigma'(x) = R(K')f(x)$ mit $R(g) = \int_{R^d} g(x)g(x)^T dx$ für eine quadratisch integrierbare Funktion g;

- $$n^{1/2}|H|^{1/4}vech(H^{1/2}(\hat{f}''(x;H) - f''(x))H^{1/2}) \xrightarrow{D} N(0, \Sigma''(x)) \tag{3.24}$$
  wobei $vech$ für den Vektor der Elemente einer oberen Dreiecksmatrix steht mit $vech \begin{bmatrix} a & b \\ c & d \end{bmatrix} = (a, b, d)^T$ und
  $$\Sigma''(x) = R(vech(K''))f(x) \tag{3.25}$$

Beim Bestimmen der signifikanten Modi wurde getestet, ob $||vech(\hat{f}''(x))||_2$ signifikant von Null verschieden ist. Der Test sieht wie folgt aus:

- $H_0 : \ ||vech(f''(x))||_2 = 0$ gegen $H_1 : \ ||vech(f''(x))||_2 \neq 0$;
- Unter H_0 gilt:
  $$(var(\hat{f}''(x)))^{-1/2}vech(\hat{f}''(x;H)) \sim N(0, I_{d^*}),$$
  wobei
  $$var(\hat{f}''(x)) = n^{-1}|H|^{-1/2}R(vech(H^{-1/2}K''H^{-1/2}))f(x)$$
  und $d^* = n(n+1)/2$;
- Die Teststatistik
  $$W_2(x) = ||(var(\hat{f}''(x)))^{-1/2}vech(\hat{f}''(x;H))||^2$$
  ist approximativ Chi-Quadrat verteilt mit $n(n+1)/2$ Freiheitsgraden;
- Man schätzt $W_2(x)$ durch
  $$\widehat{W}_2(x) = ||(\widehat{var}(\hat{f}''(x)))^{-1/2}vech(\hat{f}''(x;H))||^2,$$
  wobei
  $$\widehat{var}(\hat{f}''(x)) = n^{-1}|H|^{-1/2}R(vech(H^{-1/2}K''H^{-1/2}))\hat{f}(x;H).$$

Der Ablehnungsbereich enthält die signifikanten Modi in f und die Regionen in der Nähe von den signifikanten Modi. Dieses Resultat kann man bestätigen, indem man testet, ob $||f'(x)||$ signifikant von Null verschieden ist. Der Test sieht wie folgt aus:

- $H_0 : ||f'(x)|| = 0$ gegen $H_1 : ||f'(x)|| \neq 0$;
- Unter H_0 gilt:
$$(var(\hat{f}'(x)))^{-1/2}\hat{f}'(x;H)) \sim N(0, I_d),$$
wobei
$$var(\hat{f}'(x)) = n^{-1}|H|^{-1/2}H^{-1/2}R(K')H^{-1/2}f(x);$$
- Die Teststatistik
$$W_1(x) = ||(var(\hat{f}'(x)))^{-1/2}\hat{f}'(x;H)||^2$$
ist approximativ Chi-Quadrat verteilt mit d Freiheitsgraden;
- Man schätzt $W_1(x)$ durch
$$\widehat{W}_1(x) = ||(\widehat{var}(\hat{f}'(x)))^{-1/2}\hat{f}'(x;H)||^2,$$
wobei
$$\widehat{var}(\hat{f}'(x)) = n^{-1}|H|^{-1/2}H^{-1/2}R(K')H^{-1/2}\hat{f}(x;H).$$

Der Ablehnungsbereich dieses Tests deutet auf diejenige Regionen hin, die kein lokales Maximum oder Minimum enthalten.

Genauso wie im uni- und bivariaten Fall sind $\widehat{W}_2(x_1)$ und $\widehat{W}_2(x_2)$ bzw. $\widehat{W}_1(x_1)$ und $\widehat{W}_1(x_2)$ hoch korreliert, wenn x_1 und x_2 nah zueinander liegen. In diesem Sinne geht es hier um eine Serie von simultanen abhängigen Tests. Um das Problem zu behandeln, wurde die Methode von Hochberg (1988) in Duong et al. (2007) folgendermaßen verwendet: Seien α das Signifikanzniveau, m die Anzahl der Tests und $P_{(1)}, ..., P_{(m)}$ die geordneten fallenden P-Werte den Nullhypothesen $H_{0,(1)}, ..., H_{0,(m)}$ entsprechend, dann werden $H_{0,(1)}, ..., H_{0,(j)}$ abgelehnt, falls $P_{(j)} \leq \alpha/(m-j+1)$, also man findet hier j_{max} mit $j_{max} = argmax_j P_{(j)} \leq \alpha/(m-j-1)$, so dass $H_{0,(1)}, ..., H_{0,(j_{max})}$ abgelehnt werden. Ein anderes Problem bei diesem Test ist, dass es Regionen gibt, in denen zu wenige Daten liegen. Die sind „too sparse for inference“. Aus diesem Grund beschränkt sich der Test nur auf denjenigen Regionen, deren effektiver Stichprobenumfang $ESS \geq 5$ ist mit

$$ESS(x) = \sum_{i=1}^{n} K_H(x - X_i)/K_H(0) \tag{3.26}$$

Duong & Wand bieten ein **R** Paket **feature**, mit dem man die signifikanten Modi der Daten mit der oben vorgestellten Methode untersuchen kann. Abbildung 3.20 zeigt das Resultat aus **feature** für die Daten in Beispiel 3.1.1 (vgl. Abs. 5.1). In Abbildung 3.20 ist kein signifikanter Modus unter Signifikanzniveau 0,05 zu erkennen, was verschieden von den obigen Resultaten aus Stuetzle (2003) und Stuetzle et al. (2007) ist. Eigentlich ist die Identifizierung der signifikanten Modi in den Daten eine schwierige Aufgabe und

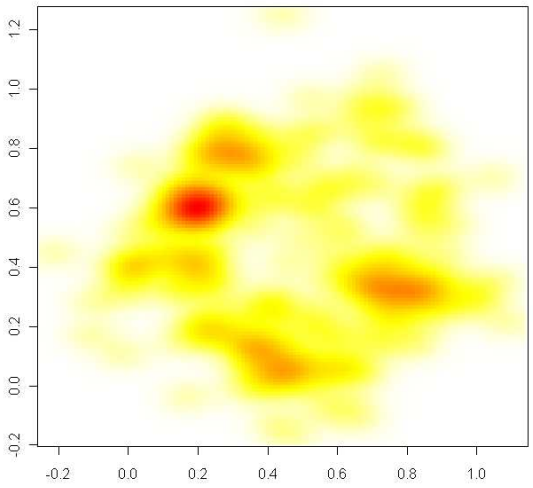

Abbildung 3.20: Resultat aus **feature** für die Daten in Beispiel 3.1.1

deswegen wurde eine Vielzahl von statistischen Methoden in den letzten Jahren dafür entwickelt. In der Regel liefern aber verschiedene Methoden unterschiedliche Resultate. Aus diesem Grund ist es nützlich, diese Resultate auf eine geeignete Art und Weise zu visualisieren und anhand der Grafiken zu vergleichen, damit man die unterschiedlichen Versionen der Datenstruktur erhalten kann. In Kapitel 4 geht man auf die grafischen Methoden weiter ein.

3.3 Dichteschätzer basiertes hierarchisches Clustering

Das Dichteschätzer basierte hierarchische Clustering ist eine Variante der Generalized Single Linkage Methode von Stuetzle et al. (2007) und wurde in Nugent (2007) vorgestellt. Das geprunte Dendrogramm aus dem Dichteschätzer basierten Single bzw. Complete Linkage Verfahren wird in dieser Arbeit als Dichteschätzer Basiertes Cluster Dendrogramm genannt, das man wie folgt konstruieren kann:

- Man berechnet einen geeigneten Dichteschätzer $\hat{f}$ anhand der empirischen Daten. In diesem Abschnitt werden ein paar Varianten vom Modell in (3.2) als $\hat{f}$ für die Konstruktion des Dichteschätzer Basierten Cluster Dendrogramms benutzt;
- Als Distanzmaß $d(x, y)$ von x und y verwendet man

$$d(x, y) = 1/min_{t\in[0,1]}\hat{f}(tx + (1 - t)y)$$

  falls $x \neq y$, und $d(x, y) = 0$, falls $x = y$;

- Ein Dichteschätzer Basiertes Dendrogramm bezieht sich auf das Dendrogramm aus dem Single bzw. Complete Linkage Clustering mit $d(x, y)$ als Distanzmaß von x und y;
- Ein Dichteschätzer Basiertes Cluster Dendrogramm ist eine geprunte Version des Dichteschätzer Basierten Dendrogramms, deren Blätter für die signifikanten Cluster stehen.

In der explorativen Datenanalyse bietet sich das Dichteschätzer Basierte Cluster Dendrogramm als eine nützliche Visualisierungsmethode an, um die unbekannte Datenstruktur insbesondere die Clusterstruktur der Daten aufzudecken. In diesem Abschnitt wird das Dichteschätzer basierte hierarchische Verfahren bzw. das entsprechende Dichteschätzer Basierte Cluster Dendrogramm anhand zwei praktischer Datensätze (**Olivenöl** (presphered) und **Italienwein** Daten) vorgestellt. Man vergleicht auch deren Resultate mit den Ergebnissen aus dem Single bzw. Complete Linkage Clustering mit $d(x, y) = ||x-y||_2$ und Kmeans-Verfahren. Kurz zu erwähnen ist, dass sich ein Single bzw. Complete Linkage Clustering in dieser Arbeit nur auf das hierarchische Verfahren mit $d(x, y) = ||x-y||_2$ bezieht, es sei denn, falls man „Dichteschätzer basiertes„ davor setzt.

Im Folgenden wird die Konstruktion eines Dichteschätzer Basierten Cluster Dendrogramms zuerst anhand eines kleinen Beispiels veranschaulicht. Der Einfachheit halber simuliert man hier einfach 11 Punkte aus einer gemischten Normalverteilung mit 4 Komponenten ohne Berücksichtigung der Ziehungswahrscheinlichkeit: 2 Punkt aus $N(1; 0, 2)$, 2 Punkte aus $N(4; 0, 2)$, 3 Punkte aus $N(2; 0, 2)$ und die anderen 4 Punkte aus $N(3; 0, 2)$. In Abbildung 3.21 zeigt man die Daten und einen Kerndichteschätzer mit Bandbreite $0, 2$. Das entsprechende Dichteschätzer Basierte Dendrogramm aus dem Single Linkage Verfahren zeigt man in Abbildung 3.22. Ein paar Erklärungen dazu:

- Die X-Achse entspricht der X-Achse in der Grafik in Abbildung 3.21;
- Die Y-Achse steht für das Dichteniveau eines Knotens im Dendrogramm;
- Die Y-Koordinaten für die Blätter im Dendrogramm sind $0, 7$;
- Die Dichteniveaus, an denen die Knoten des Dendrogramms gesplittet worden sind (vgl. Abbildung 3.23), werden in der Grafik in Abbildung 3.21 und 3.22 in horizontalen Linien gezeichnet;
- Die grüne/blaue/orange Linie verbindet zwei Einzelpunkte/einen Einzelpunkt mit einer Punktgruppe/zwei Punktgruppen;
- Die Ziffer unter jedem Nicht-Blatt-Knoten steht für das Runt Excess Maß (vgl. Abs. 3.2), d.h., das minimale Excess Maß von seinen beiden Subknoten.

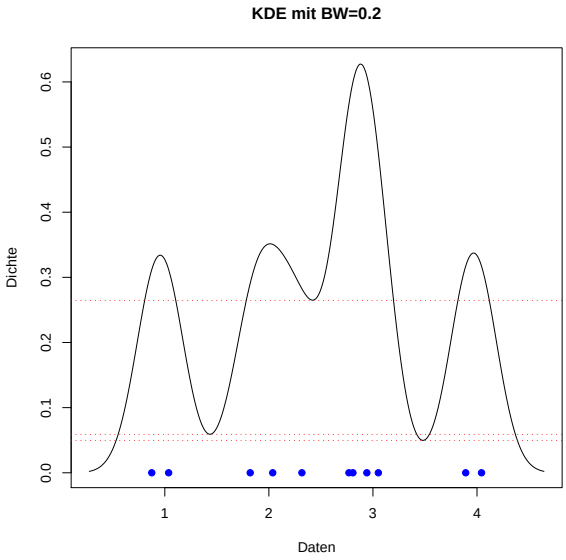

Abbildung 3.21: Simulierte 11 Punkte aus einer Normalverteilung mit 4 Komponenten und Kerndichteschätzer mit $h = 0,2$

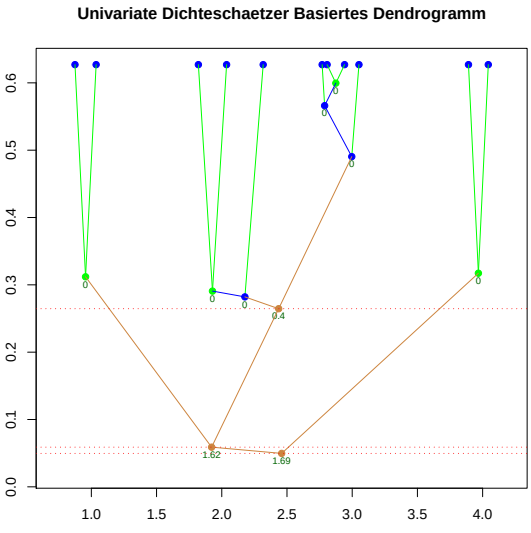

Abbildung 3.22: Dichteschätzer Basiertes Dendrogramm für die 11 simulierten Punkte

In Abbildung 3.22 ist die Clusterstruktur der Daten im Dichteschätzer Basierten Dendrogramm gut zu erkennen, weil der Umfang der Daten klein ist. In der Situation, dass der Umfang der Daten groß ist, werden nur diejenige Knoten mit großem Runt Excess Maß gesplittet, um die Datenstruktur besser zu zeigen. Somit erhält man ein Dichteschätzer Basiertes Cluster Dendrogramm. In Abbildung 3.23 stellt man ein Dichteschätzer Ba-

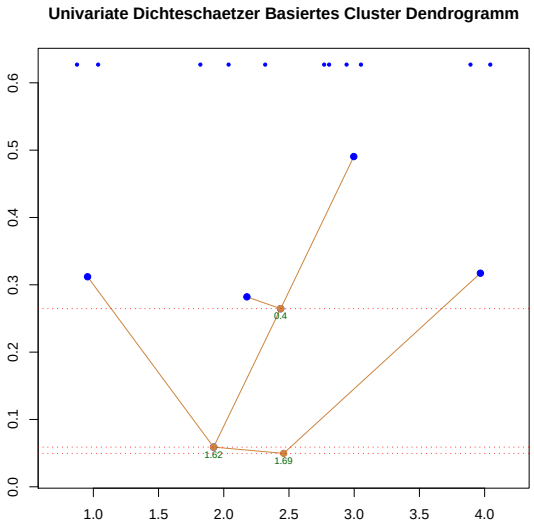

Abbildung 3.23: Dichteschätzer Basiertes Cluster Dendrogramm für die 11 simulierten Daten

sierten Cluster Dendrogramm dar. Man sieht in Abbildung 3.23, dass die 11 simulierten Daten in 4 Clustern aufgeteilt worden sind.

Aufgrund eines vorhandenen Dichteschätzer Basierten Dendrogramms werden die Daten dadurch in Clustern aufgeteilt, indem man das Dendrogramm auf eine geeignete Art und Weise abschneidet. Traditionellerweise wird ein Dendrogramm in der Regel entweder nach dem Fusionsniveau oder nach einer vorgegebenen Anzahl von Clustern abgeschnitten. Mit traditionellem Schneiden eines Dichteschätzer Basierten Dendrogramms kann man aber die Clusterstruktur der Daten nicht gut untersuchen, weil dadurch nur die High Density Cluster über einem bestimmten Dichteniveau berücksichtigt werden können, während die gesamte Clusterstruktur der Daten (über verschiedenen Dichteniveaus) meist vom Interesse ist. Man schneidet das Dichteschätzer Basierte Dendrogramm hier nach dem Runt Excess Maß eines Knotens im Dendrogramm.

Im Folgenden wird das Dichteschätzer Basierte Cluster Dendrogramm anhand der oben erwähnten zwei Beispiele veranschaulicht. Man geht wie folgt vor:

- Man verwendet hier vier Varianten von dem Modell in (3.2) als $\hat{f}$ für die Konstruieren des Dichteschätzer Basierten Dendrogramms. Die sind, V_1: Fest-Kerndichteschätzer

mit LSCV Bandbreiten bei **Olivenöl** Daten und mit Normal-Reference Bandbreiten bei **Italienwein** Daten; V_2: Fest-Kerndichteschätzer mit Normal-Reference Bandbreiten bei **Olivenöl** Daten und mit den Standardabweichungen der Variablen als Glättungsparameter bei **Italienwein** Daten; V_3: Gemischtes Modell aus dem CEM-Algorithmus (vgl. Abs. 3.1) mit V_2 als Pilot-Dichteschätzer; V_4: Gemischtes Modell aus dem Model Based Clustering von Fraley & Raftery (2002);
Eine Bemerkung dazu:
Bei **Italienwein** Daten verwendet man die Normal-Reference Bandbreiten und Standardabweichungen der Variablen als Glättungsparameter bei Kerndichteschätzung, weil der Datensatz 178 Punkte in 13 Dimensionen hat und ein Kerndichteschätzer mit kleinem Glättungsparameter in diesem Fall kein sinnvolles Ergebnis liefert.

- Man konstruiert das Dichteschätzer Basierte Single bzw. Complete Linkage Dendrogramm auf Basis von V_1, V_2, V_3, V_4. Es gibt also 8 Dendrogramme in jedem Beispiel;
- Ein Dichteschätzer Basiertes Dendrogramm wird nach dem Runt Excess Maß geprunt;
- Bei der Auswahl von einem geeigneten Runt Excess Maß für das Pruning des Dendrogramms wird das Quasi-Ellenbogen-Kriterium verwendet;
- In manchen Fällen wählt man mehrere Runt Excess Maße für das Pruning des Dendrogramms aus, wenn es Unsicherheit beim Verwenden des Quasi-Ellenbogen-Kriteriums besteht;
- Man vergleicht das Resultat aus dem Dichteschätzer basierten hierarchischen Verfahren mit den Ergebnissen aus dem Single bzw. Complete Linkage Verfahren und der Kmeans-Methode. Beim Vergleichen der obigen Resultate wird der Adjusted Rand Index benutzt;
- Die Resultate werden in erster Linie tabellarisch und grafisch dargestellt.

Beispiel 3.3.1 (**Olivenöl** Daten):

Der Datensatz enthält 572 Stichproben von Olivenöl, die aus 3 Regionen Italiens kommen. Zwei kategorielle Variablen in den Daten sind die 3 Regionen und 9 dazu zugehörige Gebiete. Stetige Variablen in den Daten sind 8 Fettsäure-Messungen. In diesem Beispiel teilt man die 572 Stichproben anhand der 8 Fettsäure-Messungen in Clustern auf und vergleicht man das Resultat mit den vorgegebenen Regionen und Gebieten. Man geht in diesem Beispiel davon aus, dass die Daten nach den 8 Fettsäure-Messungen komplett in den vorgegebenen Regionen und Gebieten aufgeteilt werden können.

Abbildung 3.24 zeigt jeweils die 20 größten Runt Excess Maße aus dem Dichteschätzer basierten Single bzw. Complete Linkage Clustering auf Basis von V_1, V_2, V_3, V_4. Wenn man beispielsweise **2S14** als die Abkürzung für das 4-Cluster Modell (**4** recht) aus dem

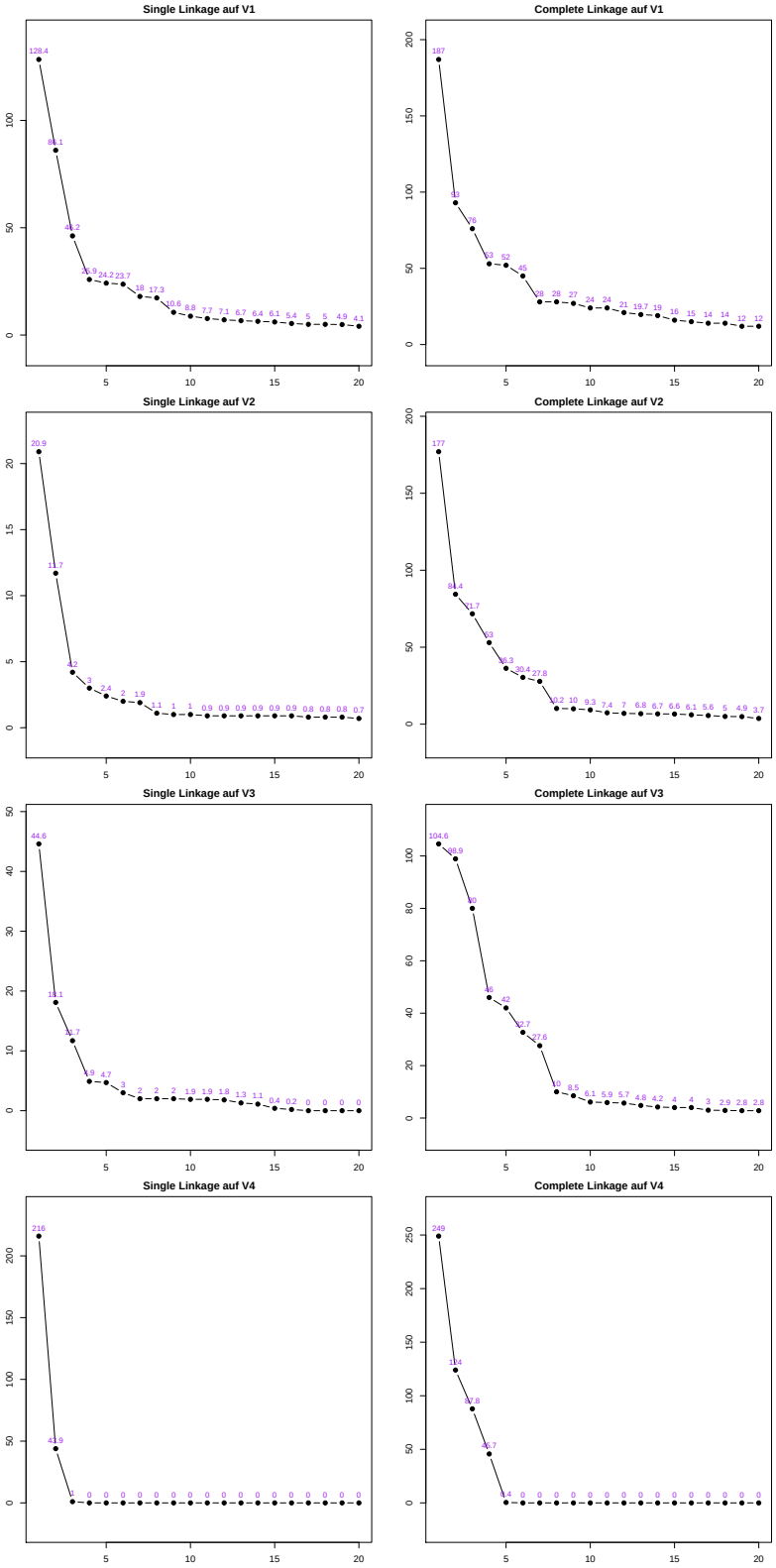

Abbildung 3.24: 20 größte Runt Excess Maße aus dem Dichteschätzer basierten Single bzw. Complete Linkage Clustering auf Basis von V_1, V_2, V_3, V_4 in Beispiel 3.3.1

ersten (**1** mittel rechts) geprunten Single Linkage Dendrogramm (**S** mittel links) auf Basis von V_2 (**2** links) benutzt, dann sieht das Pruning in diesem Beispiel wie folgt aus:

1S14 an $\epsilon = 46,2$; 1S29 an $\epsilon = 17,3$; 1C12 an $\epsilon = 187$; 1C27 an $\epsilon = 45$;
2S15 an $\epsilon = 11,7$; 2C12 an $\epsilon = 177$; 2C28 an $\epsilon = 27,8$; 3S14 an $\epsilon = 11,7$;
3C14 an $\epsilon = 80$; 3C27 an $\epsilon = 27,8$; 4S13 an $\epsilon = 43,9$; 4C15 an $\epsilon = 45,7$.

Eine kleine Bemerkung dazu:

Nach dem Quasi-Ellenbogen-Kriterium wählt man nicht immer ein optimales Runt Excess Maß für das Prunen des Dendrogramms, z.B., wenn man das Dendrogramm aus 2S15 mit Runt Excess Maß $\epsilon = 3,0$ anstatt von $\epsilon = 11,7$ (aus dem Quasi-Ellenbogen-Kriterium) prunt, dann erhält man ein 5-Cluster Modell mit Adjusted Rand Index $0,88$ und $0,53$ mit den vorgegebenen Regionen und Gebieten, was ein viel besseres Resultat darstellt.

Als Beispiel werden die Resultate aus 1S14 und 2C28 im Dichteschätzer Basierten Cluster Dendrogramm in Abbildung 3.25 dargestellt. Die Ziffer in der Grafik in Abbildung 3.25

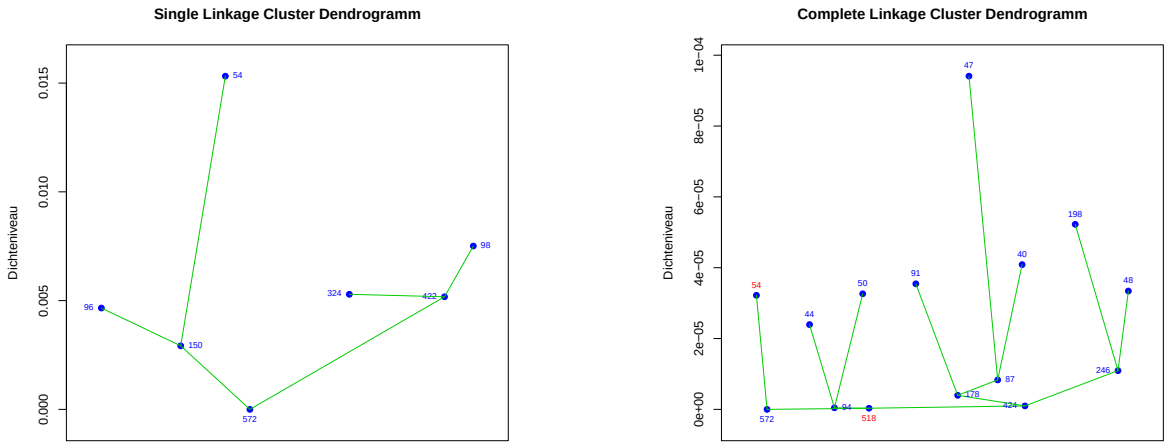

Abbildung 3.25: Resultate aus 1S14 (Grafik links) und 2C28 (Grafik rechts) im Dichteschätzer Basierten Cluster Dendrogramm in Beispiel 3.3.1

zeigt die Anzahl der Daten in dem Cluster, für den der entsprechende Knoten neben der Ziffer steht.

In Tabelle 3.2 vergleicht man die Resultate aus dem Dichteschätzer basierten hierarchischen Clustering mit denen aus dem Model Based Clustering (Fraley & Raftery 2002), Complete Linkage Clustering, Kmeans-Verfahren und den vorgegebenen Klassen (*Region* und *Area*). Dazu ein paar Erklärungen:

- Man schneidet ein Single bzw. Complete Linkage Dendrogramm nach einer vorge-

gebenen Anzahl von Clustern;

- Die Resultate aus dem Single Linkage Verfahren werden nicht berücksichtigt, weil dieses Verfahren in diesem Fall einen großen Cluster mit ein paar Einzel-Punkt-Clustern liefert;
- Nur zwei Resultate (mit 5 und 9 Clustern) aus dem Complete Linkage Verfahren werden ausgewählt, weil ein großer Cluster mit 439 Daten immer mit beim Resultat (bis zum 8-Cluster Modell) aus dem Complete Linkage Verfahren ist;
- Die anderen Abkürzungen in Tabelle 3.2 sind, Reg.:*Region*; MBC5: 5-Cluster Modell aus dem Model Based Clustering; CL5 bzw. CL9: 5- bzw. 9-Cluster Modell aus dem Complete Linkage Verfahren; KM2: 2-Cluster Modell aus dem Kmeans Verfahren und analog für KM3, KM4, KM5, KM7, KM8 und KM9;
- Diejenige Werte in Tabelle 3.2, die größer gleich $0,70$ sind, werden in Fettschrift gezeigt.

	Reg.	Area	1S14	1S29	1C12	1C27	2S15	2C12	2C28	3S14	3C14	3C27	4S13	4C15	MBC5	CL5	CL9	KM2	KM3	KM4	KM5	KM7	KM8
Area	.48																						
1S14	**.91**	.52																					
1S29	.37	.59	.41																				
1C12	.57	.28	.63	.21																			
1C27	.42	.48	.46	.48	.28																		
2S15	.30	.17	.38	.12	.09	.15																	
2C12	.55	.27	.56	.19	.39	.24	.10																
2C28	.45	**.71**	.50	.54	.29	.52	.17	.29															
3S14	.47	.25	.55	.18	.24	.22	.68	.25	.26														
3C14	.42	.54	.42	.39	.23	.37	.09	.38	.56	.18													
3C27	.46	.69	.50	.50	.30	.50	.17	.28	.69	.26	**.78**												
4S13	**.81**	.44	**.86**	.33	**.71**	.38	.35	.54	.42	.43	.34	.42											
4C15	.50	.68	.54	.46	.35	.43	.16	.28	.63	.24	.57	.68	.54										
MBC5	.55	**.72**	.57	.49	.35	.45	.15	.29	.63	.24	.60	.69	.54	**.88**									
CL5	.07	.17	.09	.10	.09	.07	.24	.00	.17	.21	.06	.16	.05	.11	.12								
CL9	.02	.28	.03	.17	.03	.14	.12	.00	.27	.10	.16	.26	.01	.24	.25	**.71**							
KM2	.44	.27	.38	.18	.09	.20	.40	.08	.26	.56	.21	.23	.26	.23	.26	.13	.15						
KM3	**.81**	.43	**.82**	.34	.56	.40	.32	.65	.42	.43	.38	.43	**.80**	.46	.47	.19	.15	.27					
KM4	.40	.64	.43	.43	.40	.38	.09	.22	.57	.13	.48	.58	.55	.68	.69	.11	.24	.16	.39				
KM5	.53	**.75**	.56	.52	.34	.48	.20	.33	.66	.26	.62	**.73**	.52	**.74**	**.78**	.19	.32	.27	.56	**.73**			
KM7	.39	.54	.43	.56	.22	.51	.13	.21	.56	.20	.43	.55	.35	.51	.53	.09	.17	.20	.37	.45	.63		
KM8	.37	.54	.36	.54	.19	.50	.09	.21	.57	.15	.44	.51	.29	.45	.48	.13	.22	.21	.34	.47	.57	**.80**	
KM9	.38	.61	.42	.53	.22	.50	.13	.22	.62	.20	.43	.56	.35	.52	.55	.13	.21	.21	.35	.50	.59	.64	.64

Tabelle 3.2: Vergleich der Resultaten aus verschiedenen Clusteringmethoden in Beispiel 3.3.1 anhand vom Adjusted Rand Index

In Tabelle 3.2 sieht man folgendes:

- Die Aufteilung der Daten in den vorgegebenen Regionen wird durch die Modelle aus 1S14, und 4S13 gut widerspiegelt;

- Die Aufteilung der Daten in den vorgegebenen Gebieten wird durch die Modelle aus 2C28, MBC5 und KM5 gut widerspiegelt;
- Die zwei am ähnlichsten Modelle in Bezug auf den Adjusted Rand Index sind das Modell aus 1S14 und *Region*;
- Das Complete Linkage Verfahren liefert in diesem Beispiel schlechtere Resultate als die aus den anderen Verfahren;
- Das Modell aus MBC5 ist sehr ähnlich wie das Modell aus 4C15;
- Das Modell aus 3C27 ist ähnlich (0,69 in Tabelle 3.2) wie das Modell aus 2C28, weil der Fest-Kerndichteschätzer bei 2C28 als Pilot-Dichteschätzer für die Konstruktion des gemischten Modells bei 3C27 verwendet wurde und man die beiden Dendrogramme nach gleichem Runt Excess Maß (27,8) geprunt hat;
- KM5 liefert ein ähnliches Resultat wie die Modelle aus 3C27, 4C15 und MBC5. Der Grund liegt darin, dass das Kriterium von der Kmeans-Methode eigentlich zu maximieren der Likelihood Funktion eines gemischten Modells mit sphärischer Kovarianz-Matrix führt.

Es besteht nun die Frage, ob die Werte in Tabelle 3.2 von **Olivenöl** Daten abhängig sind. Im Folgenden werden die obigen Modelle anhand der **Italienwein** Daten untersucht, um diese Frage zu klären.

Beispiel 3.3.2 (**Italienwein** Daten):

Der Datensatz zeigt 178 italienische Weine und 13 Merkmale des Weins, z.B., sein Gehalt an Alkohol oder die Intensität. Die 178 italienischen Weine werden nach den 13 Merkmalen in 3 Klassen eingestuft. Die Klasse des Weins ist auch in den Daten vorgegeben. In diesem Beispiel versucht man, die Weine nach den 13 Merkmalen durch Nutzung der obigen Clusteringmethoden in Clustern aufzuteilen. Man vergleicht dann die Resultate miteinander und mit der vorgegebenen Weinklasse. Analog wie in Beispiel 3.3.1 stellt man die Resultate des Vergleichs in Beispiel 3.3.2 in Tabelle 3.3 dar. In Tabelle 3.3 werden diejenige Werte, die größer gleich $0,80$ sind, in Fettschrift gezeigt.

In Tabelle 3.3 sieht man folgendes:

- Das beste Modell ist das 3-Cluster-Modell aus 2C13, das einen Adjusted Rand Index von $0,91$ mit der vorgegebenen Weinklasse hat;
- Das Dichteschätzer basierte Complete Linkage Verfahren liefert ein besseres Resultat als die anderen Verfahren;
- Die Resultate aus 3C13 und 2C13 sind ähnlich (0,85 in Tabelle 3.3);

	Kla.	1S12	1C13	2S13	2C13	3S12	3C13	4S13	4C13	MBC2	CL2	CL3	KM2
1S12	.45												
1C13	**.85**	.54											
2S13	.38	.70	.39										
2C13	**.91**	.48	**.85**	.36									
3S12	.34	.63	.35	.68	.35								
3C13	**.83**	.51	**.85**	.37	**.85**	.32							
4S13	.57	.25	.52	.28	.54	.32	.51						
4C13	.56	.46	.65	.38	.60	.37	.64	.43					
MBC2	.53	.25	.55	.20	.54	.19	.54	.35	.44				
CL2	.30	.00	.26	.00	.26	.00	.27	.31	.12	.14			
CL3	.37	.03	.35	.05	.32	.05	.32	.33	.19	.23	.49		
KM2	.37	.01	.35	.02	.33	.03	.36	.46	.21	.19	.74	.47	
KM3	.37	.06	.37	.09	.33	.10	.32	.37	.23	.23	.42	.72	.45

Tabelle 3.3: Vergleich der Resultaten aus verschiedenen Clusteringmethoden in Beispiel 3.3.2 anhand vom Adjusted Rand Index

- Die Modelle aus dem Complete Linkage Verfahren und der Kmeans Methode können die vorgegebene Weinklasse nicht gut widerspiegeln. Der Grund liegt auch darin, dass die Standardabweichungen der Variablen sehr unterschiedlich sind.

Aus Analysen der Daten in Beispiel 3.3.1 und 3.3.2 zieht man die folgenden Schlussfolgerungen:

- Ein gepruntes Dichteschätzer basiertes Single bzw. Complete Linkage Dendrogramm hilft dabei, die Clusterstruktur multivariater Daten zu untersuchen;
- Das Resultat aus dem Dichteschätzer basierten hierarchischen Verfahren hängt stark von Modellparametern des darunter liegenden Dichteschätzers ab;
- Das Resultat aus dem Dichteschätzer basierten hierarchischen Verfahren hängt auch stark davon ab, mit welchem Runt Excess Maß man das Dendrogramm prunt. Es ist für den Zweck der Datenexploration zu empfehlen, dass man ein paar Runt Excess Maße für das Pruning des Dichteschätzer Basierten Dendrogramms auswählt und die entsprechenden Resultate grafisch darstellt, um keine nützliche Information zu verlieren. Eine dazu passende grafische Methode stellt man in Abschnitt 4.3 vor;
- In beiden Beispielen liefert das Dichteschätzer basierte Complete Linkage Verfahren ein besseres Resultat als das aus dem Dichteschätzer basierten Single Linkage Verfahren in Bezug auf den Adjusted Rand Index mit den vorgegebenen Datenklassen. Das kann folgendermaßen geklärt werden: Beim Dichteschätzer basierten Complete Linkage Verfahren wird das „max/min/max“ Prinzip verwendet. Sei E die Menge aller Kanten des vollständigen Graphs $G = (V, E)$, wobei V die Menge der Daten, und weiter seien k die Anzahl der zu fusionierenden Komponenten beim hierarchischen Verfahren und $E_{a,b}$ die Menge aller Kanten zwischen Clustern a und b mit $1 \leq a, b \leq k$, dann bedeutet „max/min/max“ folgendes:

max Man gibt jeder Kante in E ein Gewicht von

$$1/min_{t\in[0,1]}\hat{f}((1-t)X_i + tX_j)$$

für $X_i, X_j \in V$, $i \neq j$;

min Für die Fusion vergleicht man in der Tat die Kanten mit maximalem Gewicht (also mit minimaler Dichte) aus $E_{a,b}$ für $1 \leq a, b \leq k$;

max Sei E_{min} die Menge der im letzten Schritt gewählten Kanten, dann wird die Kante mit maximaler Dichte aus E_{min} gewählt, damit man die zwei durch diese Kante verbundenen Cluster fusioniert.

Im Unterschied dazu wird das „max/max/max“ Prinzip beim Dichteschätzer basierten Single Linkage Verfahren verwendet. Da sich im multivariaten Fall oft zwischen den ausgewählten Dichteniveaus „max/min/max“ und „max/max/max“ ein großes Teil des Wahrscheinlichkeitsmaßes von einem Cluster befinden kann, liefert das Dichteschätzer basierte Complete Linkage Verfahren in Beispiel 3.3.1 (8-dimensional) und 3.3.2 (13-dimensional) bessere Resultate als das entsprechende Single Linkage Verfahren. Ein anderes Beispiel mit grafischer Veranschaulichung findet man in Abschnitt 4.3.

Dazu ein paar Bemerkungen:

- Der Adjusted Rand Index misst zwar die Ähnlichkeit zwei Klassifikationen der Daten, ist aber kein absolutes Kriterium. Beim Vergleichen von zwei Klassifikationen sind grafische Methoden sehr hilfreich, z.B., durch ein diagonalisiertes Mosaicplot (Fluctuationsdiagramm) kann die Verteilung der Daten in beiden Klassifizierungen gut veranschaulicht werden;

- Genauso wie bei anderen Clustermethoden wird die Identifizierung von Ausreißern beim Dichteschätzer basierten Single bzw. Complete Linkage Verfahren vernachlässigt;

- Beim Bestimmen des Distanzmaßes von zwei Punkten X_i und X_j wird numerische Methode verwendet, d.h., man nimmt k äquidistante Punkte entlang der Kante $t \cdot X_i + (1-t) \cdot X_j$ mit $t \in [0, 1]$ und vergleicht die Dichte an diesen k Punkten, um das Distanzmaß von X_i und X_j zu bestimmen;

- Der Rechenaufwand für die Bestimmung des Distanzmaßes entlang $n(n-1)/2$ (n ist der Umfang der Daten) Kanten ist teuer, falls der Datensatz groß ist. Ein nützlicher Vorschlag wäre, dass man nur die Kanten zwischen m mit $m << n$ am nächsten von X_i entfernten Daten und X_i, für $i = 1, ..., n$, berücksichtigt. In dem praktischen Beispiel im nächsten Abschnitt wird $m = 20$ ausgewählt, weil der Umfang der Daten mit $n = 16384$ groß ist;

- Ein Dichteschätzer Basiertes Dendrogramm ist eigentlich ein Baum, wenn man die Nicht-Blatt-Knoten des Dendrogramms als Kanten eines Baums $B = (V, E_B)$ betrachtet, wobei V die Menge der Daten und E_B die Menge der Kanten. Im Folgenden wird B als Dichteschätzer Basierter Baum genannt. Das Verfahren, in dem man die Knoten eines Dichteschätzer Basierten Dendrogramms splittet, bis man n Blätter hat, ist äquivalent dazu, dass man die $n-1$ Kanten des entsprechenden Dichteschätzer Basierten Baums nach der Inverse ihrer Gewichte abbricht. In diesem Sinne ist ein Dichteschätzer Basiertes Cluster Dendrogramm nicht kontinuierlich zu erweitern, wenn man das Dendrogramm nach der Größe des Runt Excess Maßes prunt;
- Man verwendet (3.18) (bzw. (3.19)), um das Excess Maß $E(N) = \int_{D(N)}(p(x) - \lambda(N))dx$ (vgl. Stuetzle et al. (2007)) eines High Density Clusters zu schätzen, wobei N der Knoten des Cluster Baums, $\lambda(N)$ das dem Knoten N entsprechende Dichteniveau und $D(N)$ die dem Knoten N entsprechende Teilmenge vom Feature Space. Falls der Umfang der Daten klein ist, dann ist $\hat{E}(N_s)$ in (3.18) kein guter Schätzer für $E(N)$.

Zum Schluss dieses Abschnitts stellt man anhand des kleinen Beispiels mit 11 simulierten Punkten im Anfang dieses Abschnitts ein intuitives Kriterium für das Pruning des Dichteschätzer basierten Dendrogramms vor. Beim Dichteschätzer basierten Clustering kann die Modalstruktur der Daten dadurch bestimmt werden, indem man den entsprechenden Dichteschätzer Basierten Baum B zerlegt. Seien B_i, $i = 1, ..., k$ die den k High Density Regions aus dem Dichteschätzer basierten Clustering entsprechenden Subbäume von B, dann wird hier das folgende intuitive Maß für das Prunen des Dichteschätzer Basierten Dendrogramms vorgeschlagen:

$$\hat{E}_h(N) = \sum_{X_p, X_q \; adjazent \in V_{B_i}} \left| \int_{X_p}^{X_q} (\hat{f} - \lambda(N))dx \right| / Dis(V_{B_i}) \tag{3.27}$$

wobei V_{B_i} die Punktmenge von B_i ist, und $Dis(V_{B_i})$ für die gesamte Länge der Kanten in B_i steht.

Erklärungen zu $\hat{E}_h(N)$:

1. Man schätzt mit $\hat{E}_h(N)$ die Höhe des High Density Clusters anstatt des Volumens $E(N)$;
2. Dabei schätzt man die Fläche der Querschnitte des High Density Clusters auf Basis der Kanten des entsprechenden Subbaums von B und dividiert diese Fläche durch die gesamte Länge der Kanten in diesem Subbaum;
3. Die Fläche des Querschnitts $\left|\int_{X_p}^{X_q}(\hat{f} - \lambda(N))dx\right|$ schätzt man durch die gesamte Fläche einer Reihe von Trapezen auf Basis der äquidistanten Punkte X_i, $i = 1, ..., s$

entlang der Kante e_{pq} mit $X_1 = X_p$ und $X_s = X_q$. Die Y-Koordinaten der 4 Eckepunkte eines Trapezes sind $\lambda, \lambda, \hat{f}(X_a), \hat{f}(X_b)$ mit a und b adjazent in $\{1, ..., s\}$. Zu bemerken ist, dass X_i und $\hat{f}(X_i)$, $i = 1, ..., s$ für die Konstruktion von B bereits vorhanden sind. Es gibt dafür keinen extra Rechenaufwand.

Im Folgenden zeigt man $\hat{E}_h(N)$ anhand des kleinen Beispiels mit 11 simulierten Punkten vom Anfang dieses Abschnitts. In Tabelle 3.4 stellt man die Kanten des entsprechenden Dichteschätzer (Kerndichteschätzer mit $h = 0,2$) Basierten Baums dar.

Erklärungen zu Tabelle 3.4:

1. In der Tabelle werden die 10 Kanten nach der Größe ihrer minimalen Dichte aufgelistet;
2. Zu C1 und C2: Eine negative Zahl $-a$, $a > 0$ steht für einen Punkt X_a und eine positive Zahl b, $b > 0$ steht für den Cluster aus der b-ten Fusion;
3. EP1 und EP2 stehen für die zwei Endepunkte einer Kante in B;
4. Dist steht für die euklidische Distanz;
5. GP1 und GP2 stehen für die zwei Endepunkte einer Kante in der grafischen Darstellung (vgl. Abbildung 3.27);
6. GDist steht für die Länge der Kante in der grafischen Darstellung (vgl. Abbildung 3.27).

KantenNr.	C1	C2	Minimale Dichte	EP1	EP2	Dist	GEP1	GEP2	GDist
1	-7	-6	0,5996	X_6	X_7	0,1339	X_6	X_7	0,1339
2	-9	1	0,5661	X_6	X_9	0,0380	X_6	X_9	0,0380
3	-8	2	0,4905	X_7	X_8	0,1102	X_7	X_8	0,1102
4	-11	-10	0,3173	X_{10}	X_{11}	0,1518	X_{10}	X_{11}	0,1518
5	-2	-1	0,3119	X_1	X_2	0,1620	X_1	X_2	0,1620
6	-4	-3	0,2907	X_3	X_4	0,2164	X_3	X_4	0,2164
7	-5	6	0,2821	X_4	X_5	0,2806	X_4	X_5	0,2806
8	7	3	0,2646	X_8	X_3	1,2311	X_5	X_9	0,4520
9	5	8	0,0589	X_6	X_2	1,7709	X_2	X_3	0,7839
10	4	9	0,0496	X_1	X_{10}	3,1686	X_1	X_{10}	3,1686

Tabelle 3.4: Dichteschätzer (Kerndichteschätzer mit $h = 0,2$) Basierter Baum der 11 simulierten Punkte vom Anfang dieses Abschnitts

Abbildung 3.26 zeigt die 11 simulierten Daten in blauen Punkten, den Kerndichteschätzer mit $h = 0,2$ in schwarzer Kurve und die Kanten des darauf basierten Maximalen Erzeugenden Baums (bezüglich der minimale Dichte entlang einer Kante) in grünen Linien. In Abbildung 3.27 werden die High Density Cluster jeweils aus dem 1-ten, 2-ten und 3-ten Splitten des Dichteschätzer Basierten Dendrogramms in unterschiedlichen Farben gezeichnet. In Abbildung 3.26 und 3.27 sieht man folgendes:

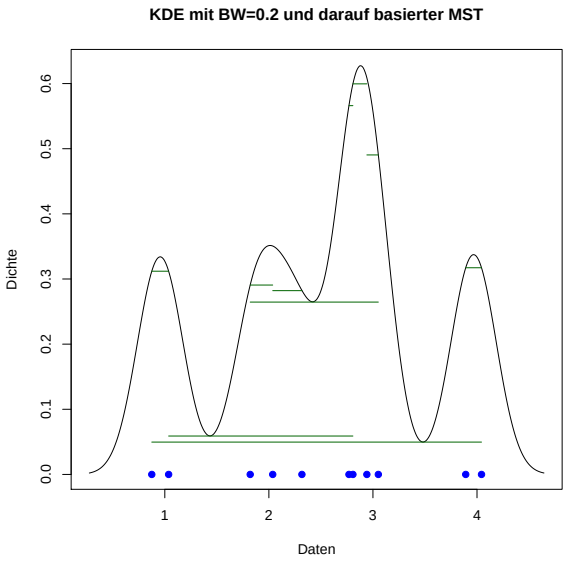

Abbildung 3.26: 11 simulierte Daten in blauen Punkten. Kerndichteschätzer mit $h = 0,2$ in schwarzer Kurve. Kanten des darauf basierten Maximalen Erzeugenden Baums in grünen Linien

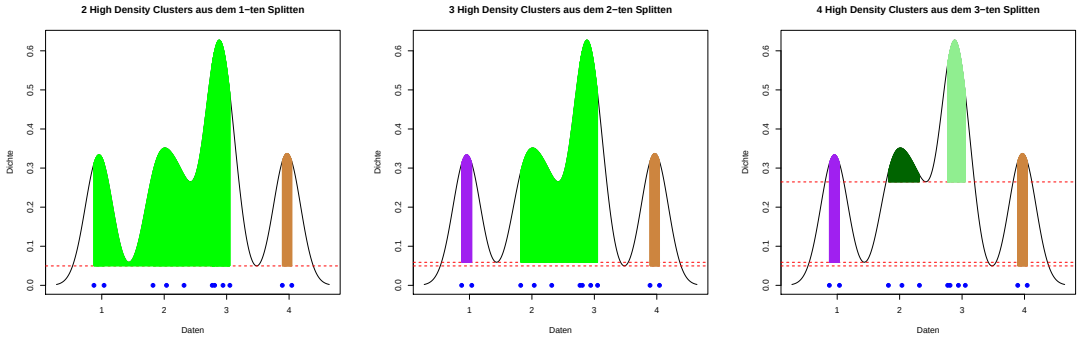

Abbildung 3.27: 2, 3 und 4 High Density Cluster aus dem 1-ten, 2-ten und 3-ten Splitten des Dichteschätzer Basierten Dendrogramms für das kleine Beispiel mit 11 simulierten Punkten

1. Eine 4-Cluster-Struktur ist klar zu erkennen;
2. Im univariaten Fall wird $E(N)$ durch $\sum_{X_p, X_q \text{ adjazent} \in V_{B_i}} \left| \int_{X_p}^{X_q} (\hat{f} - \lambda(N)) dx \right|$ approximiert, insbesondere in den High Density Clustern über einem hohen Dichteniveau;
3. Die Länge der Kante in B, die zwei Subbäume B_i und B_j von B mit $i \neq j$ verbindet, ist nicht immer die minimale euklidische Distanz zwischen V_{B_i} und V_{B_j}. In der grafischen Darstellung nimmt man die Kante mit minimaler euklidischen Distanz zwischen V_{B_i} und V_{B_j} (vgl. Tabelle 3.4), um die Überlappung der Flächen zu vermeiden. Dies führt nicht zur großen Verzerrung in der grafischen Darstellung, weil die minimale Dichte entlang dieser neuen Kante nach der Definition des Maximalen Erzeugenden Baums kleiner als die minimale Dichte aller Kanten sowohl in B_i als auch in B_j und größer als das entsprechende Dichteniveau ist.

Abbildung 3.28 zeigt die Runt Excess Maße in (3.19) (Grafik links) und die oben vorgestellten neuen intuitiven Maße (Grafik rechts) für das Prunen des Dichteschätzer Basierten Dendrogramms. Im Vergleich zu der linken Grafik in Abbildung 3.28 weist die rechte Grafik klar auf $\hat{E}_h(N) = 0,3$ aus dem Quasi-Ellenbogen-Kriterium für das Prunen des Dichteschätzer Basierten Dendrogramms hin, was zu einem 4-Cluster-Modell führt, das der Verteilung bei der Simulation entspricht. Zu bemerken ist, dass man mit $\hat{E}_h(N)$ in vielen Fällen das gleiche Resultat wie bei Nutzung des geschätzten Excess Maßes in (3.19) bekommt, was zu erwarten ist. In Abschnitt 4.3 stellt man auf Basis der Grafiken in Abbildung 3.27 eine grafische Methode vor, um die Form der High Density Cluster zu beschreiben.

3.4 Eine praktische Anwendung

In diesem Abschnitt wird die Anwendung des Dichteschätzer basierten hierarchischen Verfahrens bei Analyse großer Datensätze anhand eines praktischen Beispiels im Bereich von Fernerkundung gezeigt. Weitere Analyse ist nötig. Zuerst stellt man den **out5d** Datensatz vor.

Beispiel 3.4.1 out5d Daten

Die **out5d** Daten enthalten Messungen aus SPOT, von magnetischer Strahlung und von Radiometrics Messungen (Kalium, Thor und Uran) in 128×128 Regionen in Western Australien. Die Namen der Variablen im Datensatz sind $SPOT$, Mag, $Potas$, $Thor$ und $Uran$. Das Ziel der Datenanalyse besteht in der Untersuchung der Verteilung von Kalium, Thor und Uran in Western Australien. Die Information über *out5d* Daten befindet sich unter http://davis.wpi.edu/xmdv/datasets/out5d.html.

Abbildung 3.29 zeigt die Histogramme von Mag, $Potas$, $Thor$ und $Uran$ mit Binbreite $= 2$ in Software **Mondrian**, wobei die Daten mit $Potas \leq 164$ in rot markiert sind. Die Grafik unten in Abbildung 3.29 zeigt die 128×128 Regionen in der Landkarte. Durch

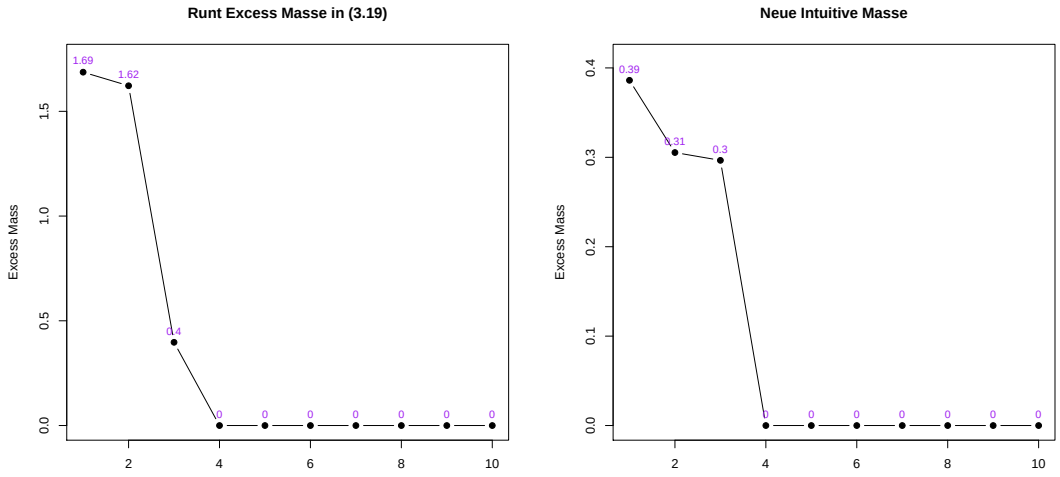

Abbildung 3.28: Runt Excess Maße in (3.19) (Grafik links) und neue intuitiven Maße (Grafik rechts) für das Prunen des Dichteschätzer Basierten Dendrogramms für das kleine Beispiel mit 11 simulierten Punkten

das Highlighting in Abbildung 3.29 werden die 128×128 Regionen grob in zwei Gruppen aufgeteilt. Die in rot markierte Gruppe besteht aus den Regionen, die in erster Linie wenig Kalium und auch relativ wenig Thor und Uran enthalten. Eigentlich kann man viel Information über die Struktur der **out5d** Daten erhalten, wenn man die interaktiven Visualisierungsmethoden von **Mondrian** ausnutzt.

Im Folgenden wird das Dichteschätzer basierte Single Linkage Verfahren auf Basis vom Kerndichteschätzer mit der Biweight Kernfunktion und vom Nearest Neighbour Dichteschätzer in die Datenanalyse mit einbezogen, um die Daten aus einem anderen Sichtwinkel zu betrachten. Man geht wie folgt vor:

- Man schätzt die unbekannte Dichte der Verteilung von Kalium, Thor und Uran und identifiziert die „High Density Regions“ in der geschätzten Dichte;
- Man untersucht die Verteilungen von Kalium, Thor und Uran in den „High Density Regions“.

Single Linkage Verfahren auf Basis vom Kerndichteschätzer

Da es bei Kerndichteschätzung das Randproblem bei *Potas* (vgl. Abbildung 3.29) gibt, wird hier ein Kerndichteschätzer wie folgt verwendet:

$$\hat{f}_r(x) = \frac{1}{n \prod_{s=1}^{d} h_s} \sum_{i=1}^{n} \prod_{s=1}^{d} K_d \left(\frac{x_s - X_{is}}{h_s} \right) \tag{3.28}$$

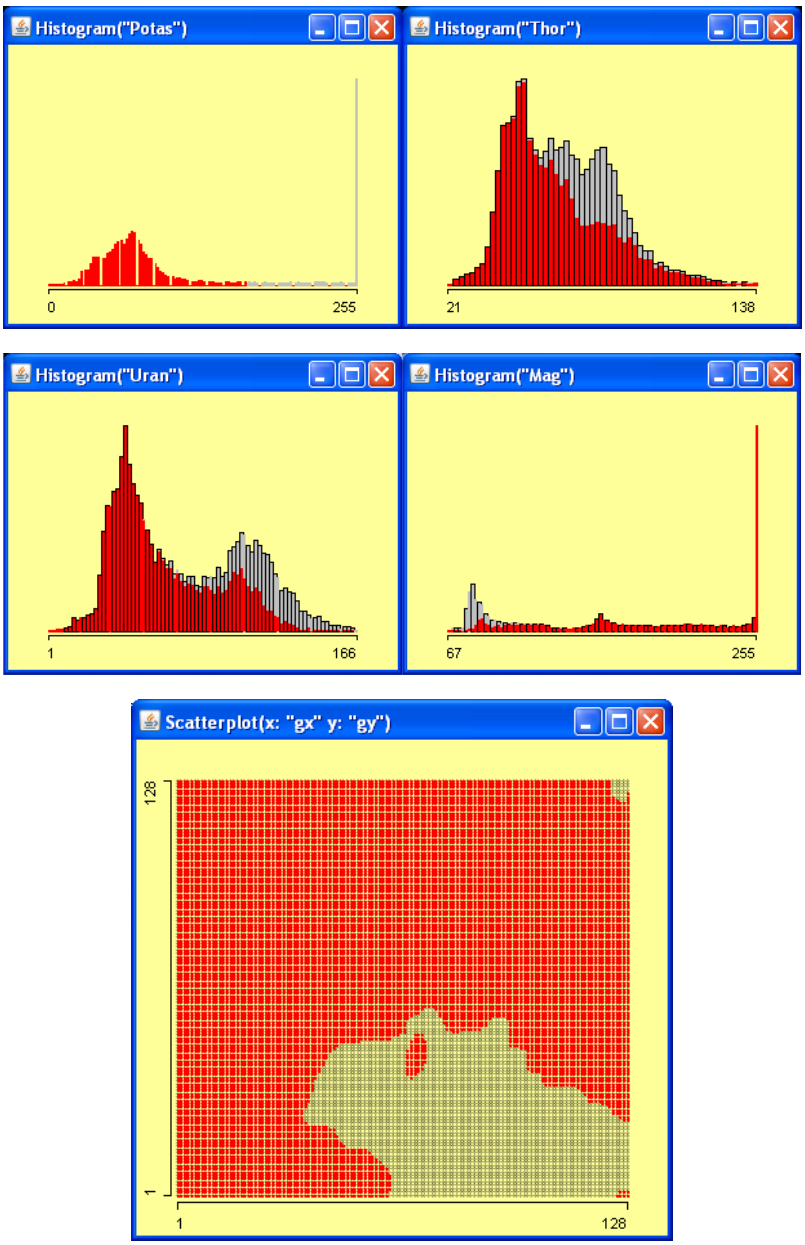

Abbildung 3.29: *Mag*, *Potas*, *Thor* und *Uran* im Histogramm in **Mondrian**. Daten mit $Potas \leq 164$ in rot markiert. Die Grafik unten zeigt die 128×128 Regionen in der Landkarte.

wobei $n = 16384$, $d = 3$, x_s, $s = 1, 2, 3$ jeweils für *Potas*, *Thor* und *Uran* steht,

$$K_1(t) = \frac{3}{4}(1 - c + \frac{5}{4}(2c - 1)(t - c)^2)(t - c + 2)^2 I_{[c-2,c]}(t) \tag{3.29}$$

für $x_1 \in (255 - h_1; 255]$ mit $c = \frac{255-x_1}{h_1}$ und

$$K_1(t) = K_2(t) = K_3(t) = \frac{15}{16}(1 - t^2)^2 I_{[-1,1]}(t)$$

für $x_1 \in [0; 255 - h_1]$.

Ein anderes Problem für die Nutzung vom Single Linkage Verfahren auf Basis von $\hat{f}_r$ liegt darin, dass der Rechenaufwand für die Bestimmung der minimalen Dichten entlang $n(n-1)/2$ Kanten wegen des Umfangs der **out5d** Daten ziemlich groß ist. Ein nützlicher Vorschlag wäre, dass man nur die Kanten zwischen m mit $m << n$ am nächsten von X_i entfernten Daten und X_i, für $i = 1, ..., n$, berücksichtigt. Unserer Erfahrung nach reicht es, einen m mit $m \in [\frac{n}{100}, \frac{n}{20}]$ zu wählen, wenn n groß ist, um einen zusammenhängenden Graph zu konstruieren. Hier in unserem Beispiel wurde $m = 20$ gewählt. Abbildung 3.30 zeigt die nach $\hat{f}_r$ eingefärbten Regionen in der Landkarte, wobei die **R** Farbpalette

```
palette(gray((0:255)/255))
```

verwendet wurde. Abbildung 3.31 zeigt die Verteilung der Farben in der Landkarte in

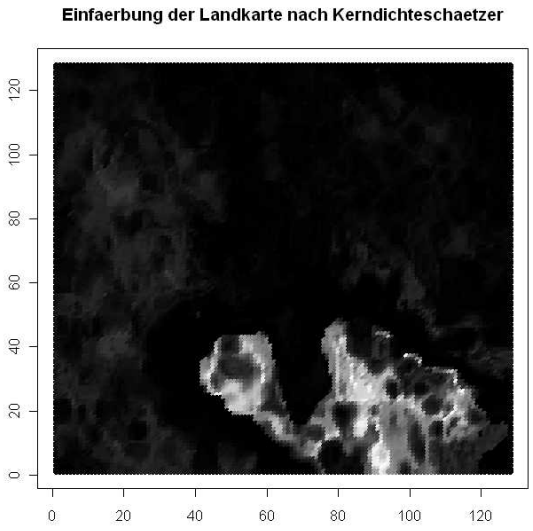

Abbildung 3.30: Einfärbung der Regionen nach $\hat{f}_r$ in der Landkarte in Beispiel 3.4.1

Abbildung 3.30 in einem selbstgewichteten Histogramm, in dem die Fläche des k-ten

Rechtecks proportional zu $\sum_{i \in X_{[k]}} |X_i|$, wobei $X_{[k]}$ für die Subgruppe der Daten im k-ten Rechteck steht, dargestellt wird. Die Grafik links in Abbildung 3.32 zeigt die 20

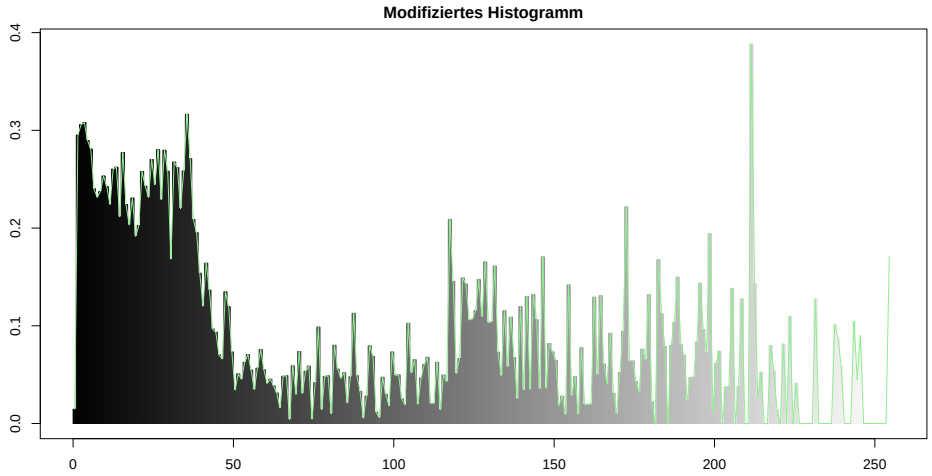

Abbildung 3.31: Verteilung der Farben in der Landkarte in Abbildung 3.30 in Beispiel 3.4.1

größten Runt Excess Maße aus dem auf $\hat{f}_r$ basierten Single Linkage Dendrogramm. Man wählt Runt Excess Maß $= 2769,4$ für das Pruning des Dendrogramms und stellt das Resultat im Visualisierungsbaum (vgl. Abs. 4.3) in der rechten Grafik in Abbildung 3.32 dar. Beim Pruning mit Runt Excess Maß $= 2769,4$ erhält man zwei Cluster mit 15577 Punkten in den High Density Regions und 807 Punkten als Noise. In Abbildung 3.33 stellt man die Verteilung von Kalium, Thor und Uran in beiden High Density Regions in Histogramm dar. Man sieht in Abbildung 3.33, dass die Aufteilung der Daten aus dem auf $\hat{f}_r$ basierten Single Linkage Verfahren sehr ähnlich wie die in den Grafiken in Abbildung 3.29 gezeigte Aufteilung ist. Das Dichteschätzer basierte Single Linkage Verfahren bietet hier eine gute theoretische Unterstützung für diese Aufteilung der Daten an.

Single Linkage Verfahren auf Basis von Nearest Neighbour Dichteschätzer

Hier wird das Runt Pruning Verfahren von Stuetzle (2003) verwendet. Die Grafik links in Abbildung 3.34 zeigt die 20 größten Runt Sizes aus dem auf $\hat{f}_{nn}$ basierten Single Linkage Dendrogramm. Man wählt Runt Size $= 910$ für das Pruning des Dendrogramms aus und stellt den entsprechenden Cluster Baum in der rechten Grafik in Abbildung 3.34 dar, in der die 8 Knoten des Cluster Baums unterschiedlich eingefärbt worden sind, damit die Verteilungen von Kalium, Thor und Uran in den entsprechenden Subgruppen der Daten im Histogramm in Abbildung 3.35 gut zu unterscheiden sind. In Abbildung 3.34 und 3.35 sieht man folgendes:

- Die Verteilungen von Kalium, Thor und Uran in den 4 Subgruppen der Daten, die

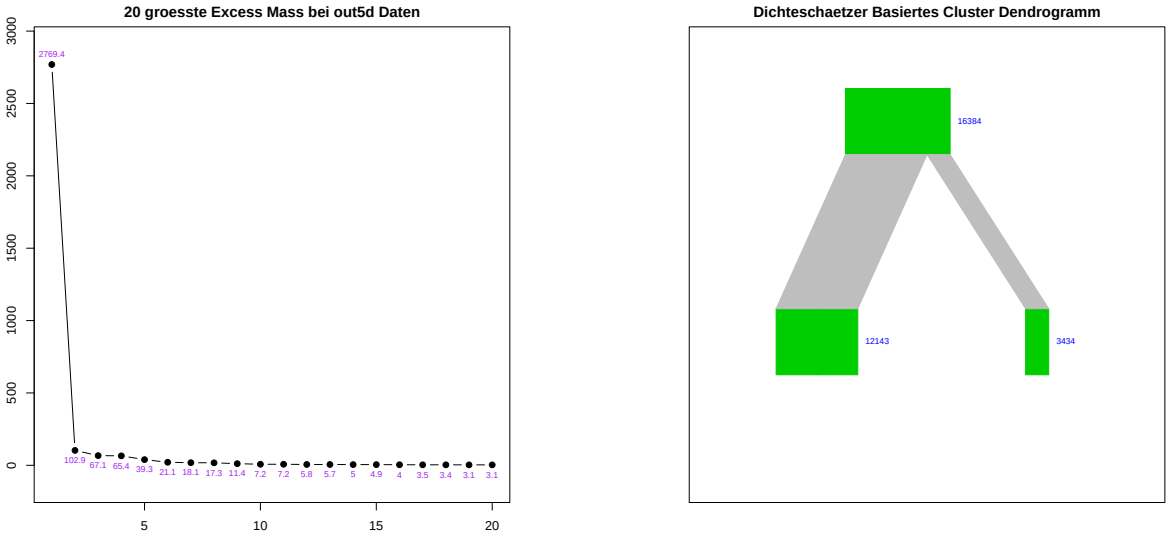

Abbildung 3.32: 20 größte Runt Excess Maße aus dem auf $\hat{f}_r$ basierten Single Linkage Dendrogramm (Grafik links). Resultat mit Runt Excess Maß $= 2769,4$ im Visualisierungsbaum (Grafik rechts)

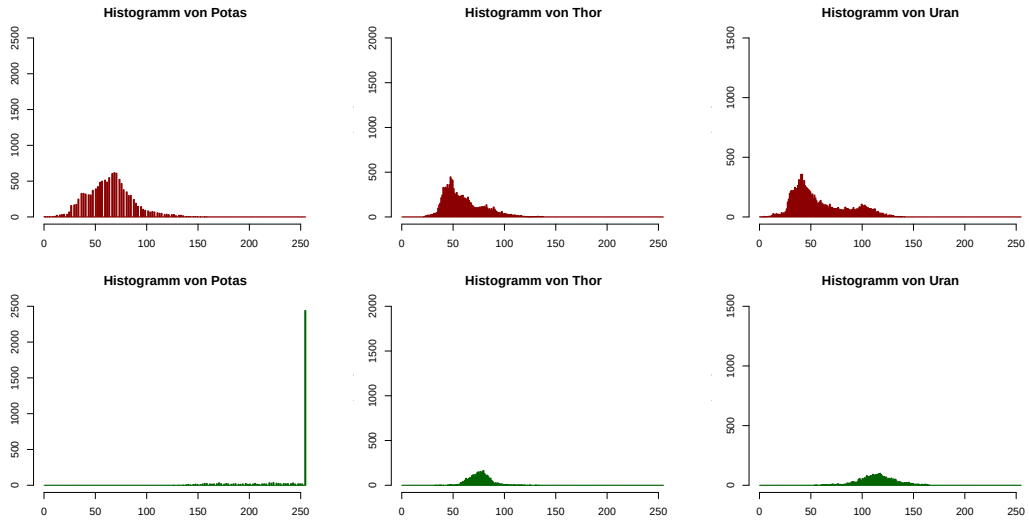

Abbildung 3.33: Verteilung von Kalium, Thor und Uran in der 1-ten High Density Region (Grafiken oben) und in der 2-ten High Density Region (Grafiken unten)

den 4 linken Knoten des Cluster Baums in Abbildung 3.34 entsprechen, verändern sich kaum;

- Die kleinere Gruppe aus dem ersten Splitten des Cluster Baums besteht aus den Regionen, die mehr Kalium und relativ mehr Uran enthalten im Vergleich zu der größeren Gruppe;
- Die kleinere Gruppe aus dem zweiten Splitten des Cluster Baums besteht aus den Regionen, die relativ weniger Uran enthalten im Vergleich zu der größeren Gruppe;
- Die kleinere Gruppe aus dem dritten Splitten des Cluster Baums besteht aus den Regionen, die relativ weniger Kalium enthalten im Vergleich zu der größeren Gruppe;
- Die kleinere Gruppe aus dem vierten Splitten des Cluster Baums besteht aus den Regionen, die relativ weniger Kalium und Uran enthalten im Vergleich zu der größeren Gruppe.

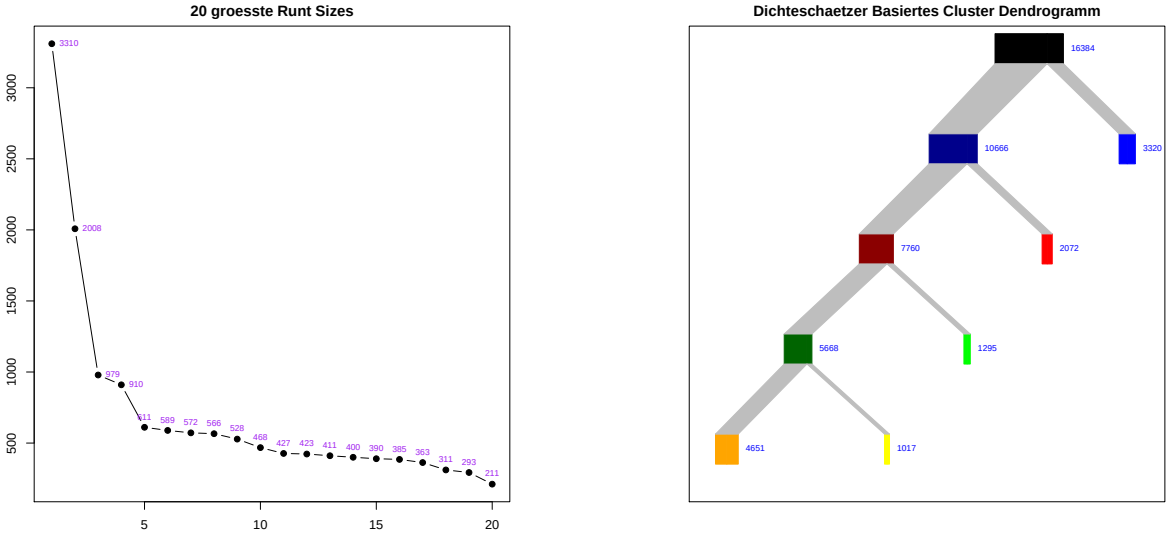

Abbildung 3.34: 20 größte Runt Sizes aus dem auf $\hat{f}_{nn}$ basierten Single Linkage Dendrogramm (Grafik links). Resultat mit Runt Size = 910 im Visualisierungsbaum (Grafik rechts)

In Hinsicht der Untersuchung der Verteilung von Kalium, Thor und Uran macht eigentlich nur das erste Splitten des Cluster Baums aus dem obigen Runt Pruning Verfahren einen Sinn. In der Tat ist das Resultat aus dem ersten Splitten des Cluster Baums aus dem Runt Pruning Verfahren sehr ähnlich wie das obige Resultat aus dem Pruning des auf $\hat{f}_r$ basierten Single Linkage Dendrogramms, wie es im Fluctuationsdiagramm der beiden Resultate in Abbildung 3.36 zu erkennen ist. In Abbildung 3.37 werden die High

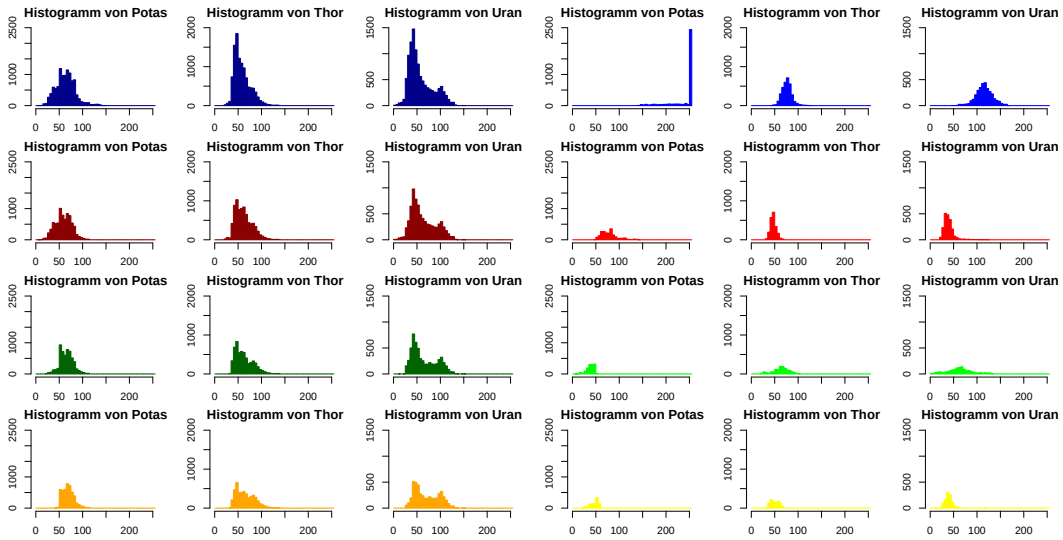

Abbildung 3.35: Verteilung von Kalium, Thor und Uran den Knoten des Cluster Baums in Abbildung 3.34 entsprechend. Die 4 Reihen der 3 linken/rechten Spalten stehen für die Verteilung von Kalium, Thor und Uran in den 4 Subgruppen der Daten, die den 4 linken/rechten Knoten des Cluster Baums entsprechen.

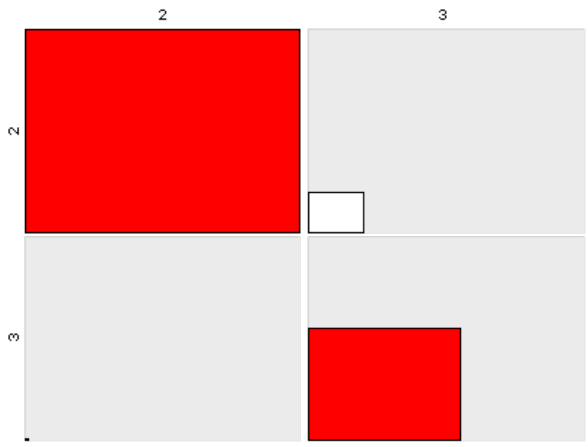

Abbildung 3.36: Fluctuationsdiagramm der beiden Resultate aus den auf $\hat{f}_r$ und $\hat{f}_{nn}$ basierten Single Linkage Verfahren in Beispiel 3.4.1

Density Regions aus den obigen beiden Verfahren in der Landkarte gezeichnet, in der die Regionen nach *SPOT* eingefärbt sind. Die Abkürzung KDBSL bzw. NNDBSL in Abbildung 3.37 steht für das Kerndichteschätzer bzw. Nearest Neighbour Dichteschätzer basierte Single Linkage Verfahren. Die Grafik in Abbildung 3.37 bestätigt die Ähnlichkeit der obigen beiden Resultate, weil es kaum Unterschied zwischen den Landkarten aus KDBSL und NNDBSL gibt.

Aus der Analyse der **out5d** Daten sind folgende Schlüsse zu ziehen:

- Bei der Untersuchung der Verteilung von Kalium, Thor und Uran ist das Dichteschätzer basierte hierarchische Verfahren hilfreich;
- Das Runt Pruning Verfahren von Stuetzle (2003) ist für die Analyse großer Datensätze gut geeignet;
- Beim Anwenden des Kerndichteschätzer basierten hierarchischen Verfahrens auf großen Datensätzen braucht man nur die Gewichte von $m \cdot n/2$ Kanten für die Konstruktion des Dichteschätzer Basierten Baums zu bestimmen, indem man jeweils die m am nächsten entfernten Nachbarn der Daten berücksichtigt;
- Das Randproblem soll in der Datenanalyse beachtet werden. In unserem Beispiel haben die Randwerte von Variable *Potas* einen großen Einfluss auf das Resultat der Datenanalyse. Deswegen liefern die beiden Dichteschätzer basierten Verfahren fast das gleiche Resultat. Es wäre vielleicht sinnvoll, nur das Teil der Daten mit $Potas < 255$ zu untersuchen. Um zu wissen, ob dies erlaubt ist, ist die Meinung eines Fachmanns aus dem Bereich der Fernerkundung erforderlich.

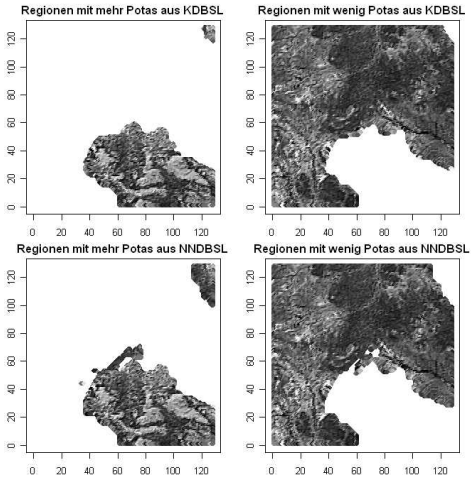

Abbildung 3.37: Resultate aus den auf $\hat{f}_r$ und $\hat{f}_n$ basierten Single Linkage Verfahren in der Landkarte in Beispiel 3.4.1

4 Dichteschätzung und Visualisierung

Visualisierung ist ein großes wissenschaftliches Gebiet und es gibt stets Diskussionen über die Subgebiete der Visualisierung und deren genaue Definitionen. In diesem Kapitel beschränken wir uns auf die Visualisierung in der explorativen statistischen Datenanalyse, die in erster Linie zur Datenexploration dient. Die Hauptaufgabe der Visualisierung in der explorativen Datenanalyse liegt darin, die Eigenschaften der Daten wie Muster, Cluster, Korrelation usw. aufzudecken.

Der Schwerpunkt dieses Kapitels liegt in der Visualisierung der multivariaten Daten. Heutzutage kommen in zahlreichen Wissenschaftsbereichen sowohl der Naturwissenschaft als auch der Sozialwissenschaft oft empirische Daten mit vielen Merkmalen vor, was impliziert, dass die Daten im statistischen Sinne im hochdimensionalen Raum verteilt sind. In den letzten Jahren wurden verschiedene Methoden vorgeschlagen, mit denen man multivariate Daten grafisch darstellen kann. Eine interessante Zusammenfassung dieser Methoden findet man in Chen, Härdle und Unwin (2008).

Das vorliegende Kapitel gliedert sich in 3 Teile. Zuerst werden ein paar übliche Visualisierungsmethoden für die multivariaten Daten vorgestellt. Dann im zweiten Teil wird gezeigt, wie man die Information aus Dichteschätzung mit in die Datenvisualisierung einbezieht. Schließlich werden zwei Visualisierungsmethoden vorgeschlagen, mit denen man die Information aus dem Dichteschätzer basierten hierarchischen Verfahren grafisch darstellen kann.

4.1 Vorstellung der üblichen Visualisierungsmethoden

In diesem Abschnitt werden die folgenden Visualisierungsmethoden anhand simulierter Daten vorgestellt: Scatterplot Matrix, Parallel Koordinaten Plot (Inselberg (1985)), Andrews Kurve (Andrews (1972)), Glyph-Visualisierung auf Basis von statistischer Software **Gauguin** (Gribov & Unwin (2006)), Heatmap und Projection Pursuit auf Basis von statistischer Software **GGobi** (Swayne et al. (1999)). Das Ziel dieses Abschnitts besteht darin, einen ersten Blick in die multivariate Datenvisualisierung zu geben. Zuerst stellt man in Beispiel 4.1.1 die simulierten Daten vor. Für

Beispiel 4.1.1 (Simulierte Daten)

Man simuliert hier 500 Daten $X_1, ..., X_{500} \in R^5$ aus einer gemischten Normalverteilung mit Dichtefunktion

$$f_{b1} = \sum_{i=1}^{6} \phi(\mu_i, \Sigma_i)/6, \tag{4.1}$$

wobei $\mu_1 = (-1,5; 1,5; 1,5; 1,5; 1,5)^T$, $\mu_2 = (1,5; -1,5; 1,5; 1,5; 1,5)^T$,
$\mu_3 = (1,5; 1,5; -1,5; 1,5; 1,5)^T$, $\mu_4 = (1,5; 1,5; 1,5; -1,5; 1,5)^T$,
$\mu_5 = (1,5; 1,5; 1,5; 1,5; -1,5)^T$, $\mu_6 = (0,0,0,0,0)^T$ und $\Sigma_i = I_5$ für $i = 1, ..., 6$.

In f_{b1} sieht man, dass das gemischte Modell aus 6 Komponenten besteht und die 6 Komponenten klar voneinander getrennt sind. Im Folgenden werden die oben erwähnten multivariaten Visualisierungsmethoden anhand des Beispiels 4.1.1 vorgestellt. Es ist interessant zu wissen, ob die 6 Modi in den simulierten Daten durch Visualisierung zu erkennen sind.

Scatterplot Matrix

Seien $Y_1, ..., Y_d$ Zufallsvariablen aus R^d, dann werden in einer Scatterplot Matrix die Beobachtungen für jeweils zwei Zufallsvariablen Y_i und Y_j mit $i, j \leq d$, $i \neq j$ in einem Streudiagramm dargestellt. In diesem Sinne geht es bei einer Scatterplot Matrix eigentlich um $d(d-1)/2$ marginale Verteilungen. Abbildung 4.1 zeigt die Scatterplot Matrix für die Daten in Beispiel 4.1.1, wobei die Daten aus verschiedenen Komponenten unterschiedlich eingefärbt sind. In Abbildung 4.2 fügt man die geschätzten Dichtefunktionen (Kerndichteschätzer mit LSCV-Bandbreiten) der 2D marginalen Verteilungen in roten Contour Linien zu den Scatterplots hinzu. Man sieht in Abbildung 4.1 und Abbildung 4.2, dass die 6 Modi in den Daten durch die Scatterplot Matrix nicht widerspiegelt werden können. In der Scatterplot Matrix in Abbildung 4.2 stellt man eigentlich auch die Information aus Kerndichteschätzung dar durch Hinzufügung der Contour Linien zu den Scatterplots.

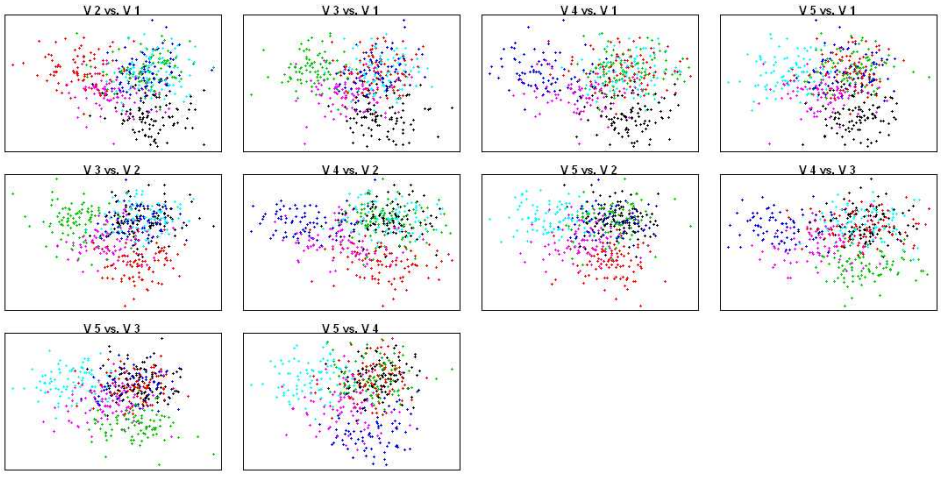

Abbildung 4.1: Daten in Beispiel 4.1.1 in der Scatterplot Matrix

Parallel Koordinaten Plot

Bei der Parallelkoordinaten-Darstellung definiert man eine Abbildung von $R^d \to R^2$, indem die d Achsen parallel nebeneinander senkrecht zur Abszisse in einer 2 dimensionalen

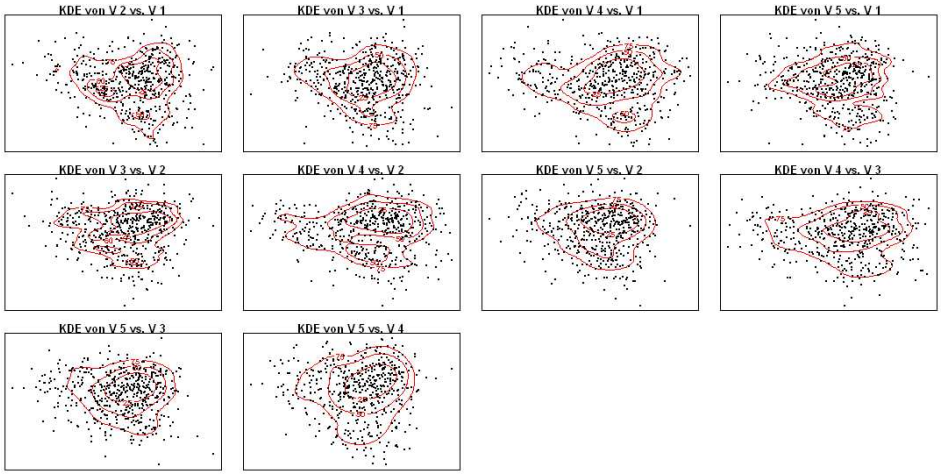

Abbildung 4.2: Daten in Beispiel 4.1.1 (mit Hinzufügung des Kerndichteschätzers mit LSCV-Bandbreiten in roten Contour Linien zum Scatterplot) in der Scatterplot Matrix

Ebene dargestellt werden. Die Daten vom d dimensionalen euklidischen Raum werden in der Tat auf diese Art und Weise in die 2 dimensionale Ebene projiziert. Der Parallel Koordinaten Plot bietet sich als nützliches Werkzeug für die Visualisierung multivariater Daten an, mit dem man die in den multivariaten Daten versteckten Strukturen wie z.B. Korrelation und Clusterstruktur untersuchen kann. Für eine ausführliche Beschreibung dieser Methode verweisen wir auf die Arbeit von Inselberg (1985, 2006). Die linke Grafik in Abbildung 4.3 zeigt die Daten von Beispiel 4.1.1 im Parallel Koordinaten Plot, wobei die Daten aus 6 Komponenten unterschiedlich eingefärbt sind. Die Parallelkoordinaten-Darstellung hat einen Nachteil, nämlich, dass die Datenstruktur wegen der Überlappung der Linien nicht gut erkennbar sein kann, wenn der Umfang der Daten groß ist. Im Fall eines großen Datensatzes kann man z.B. durch Veränderung der Reihenfolge der Koordinaten oder durch Verwendung von anderen grafischen Mitteln wie z.B. durch Einbezug von Alpha Blending in die grafische Darstellung die Erkennbarkeit der Datenstruktur verbessern. In der rechten Grafik in Abbildung 4.3 werden die Daten im Parallel Koordinaten Plot mit Alpha Blending $= 0,2$ dargestellt. Man sieht in Abbildung 4.3, dass die 6-Komponenten Struktur durch die Parallelkoordinaten-Darstellung nicht direkt erkennbar ist. Für den Zweck der Daten Exploration wird ein Parallel Koordinaten Plot in der Praxis oft interaktiv dargestellt, die üblicherweise durch Java-Programmierung verwirklicht wird. Die Software **Mondrian** aus ROSUDA der Universität Augsburg bietet z.B. eine volle interaktive Darstellung des Parallel Koordinaten Plots. Nähere Information befindet sich unter http://rosuda.org/software/Mondrian/Mondrian.html.

Andrews Kurve

Bei der Andrews Kurve definiert man eben eine Abbildung von $R^d \to R^2$, indem man

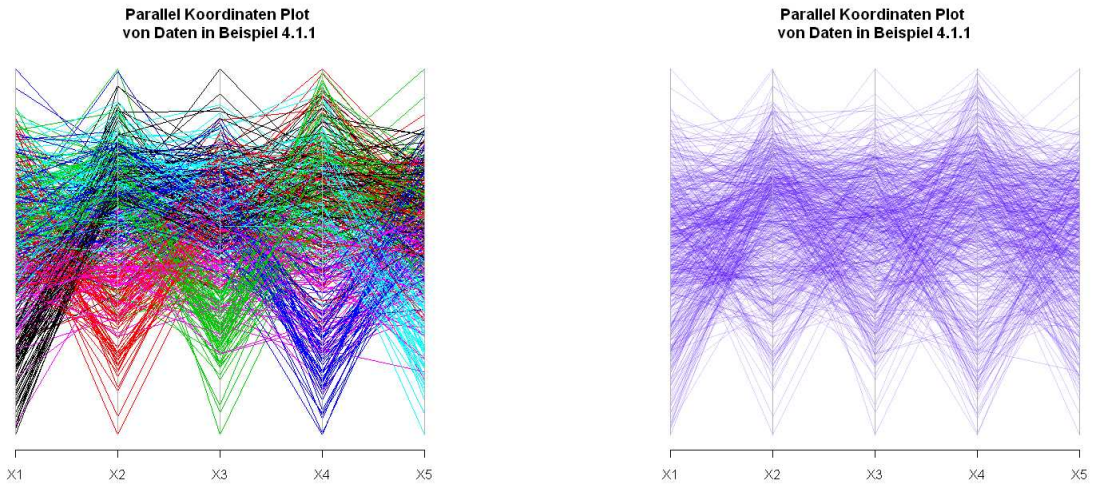

Abbildung 4.3: Daten in Beispiel 4.1.1 im Parallel Koordinaten Plot. Die Linien aus den 6 Komponenten werden unterschiedlich eingefärbt (Grafik links). Mit Einbezug von Alpha Blending $= 0,2$ (Grafik rechts)

einen Punkt $X_i = (X_{i1}, ..., X_{id})^T$, $i \in \{1, ..., n\}$ in R^d durch die folgende Funktion

$$f(t) = \frac{X_{i1}}{\sqrt{2}} + X_{i2} * sin(t) + X_{i3} * cos(t) + X_{i4} * sin(2t) + X_{i5} * cos(2t) + ... \qquad (4.2)$$

in eine Kurve in R^2 transformiert. In Abbildung 4.4 werden die Daten in Beispiel 4.1.1 in die Andrews Kurve ohne Alpha Blending (Grafik links) und mit Alpha Blending $= 0,2$ (Grafik rechts) dargestellt. Man sieht in der rechten Grafik in Abbildung 4.4, dass die wahre Datenstruktur in der Andrews Kurve genauso wie beim Parallel Koordinaten Plot nicht direkt erkennbar ist.

Glyph-Visualisierung

Bei der Glyph-Visualisierung definiert man ein visuelles Mapping von Variablen $\rightarrow$ Elementen der Grafik, was bedeutet, dass die Parameter (Variablen) der Daten in systematischer Weise zu den grafischen Elementen zugeordnet werden. Die grafischen Elemente sind z.B. Position, Länge, Breite, Form, Farbe, Größe, Orientierung, Texture einer Grafik. Dieses visuelle Mapping besteht aus

- $1 \rightarrow 1$ Mapping, d.h., verschiedene Variablen werden zu unterschiedlichen grafischen Elementen zugeordnet;
- $1 \rightarrow n$ Mapping, d.h., eine Variable wird zu mehreren grafischen Elementen zugeordnet;

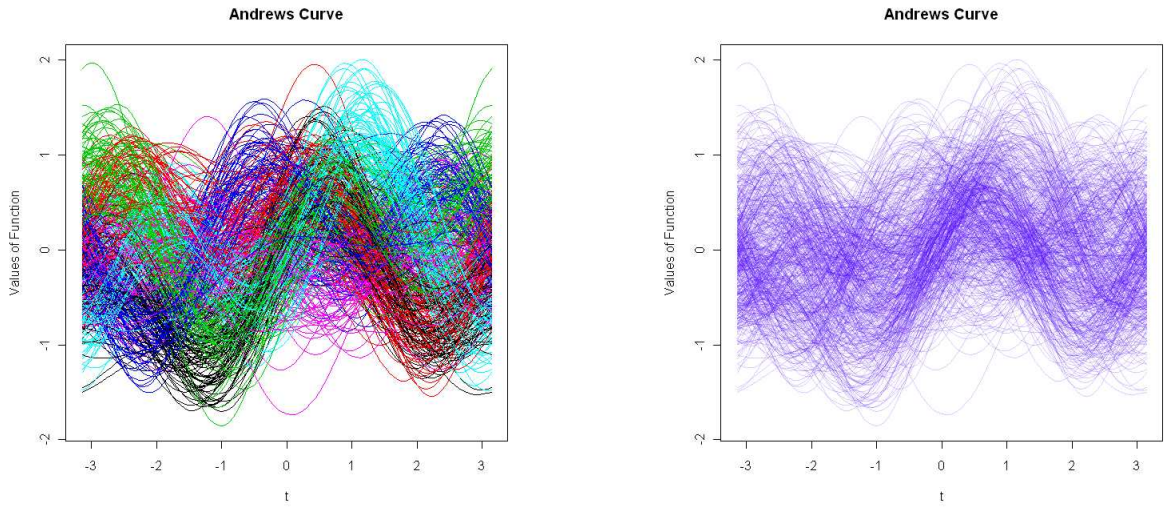

Abbildung 4.4: Daten in Beispiel 4.1.1 in der Andrews Kurve. Die Linien aus den 6 Komponenten werden unterschiedlich eingefärbt (Grafik links). Mit Einbezug von Alpha Blending $= 0,2$ (Grafik rechts)

- $n \rightarrow 1$ Mapping, d.h., mehrere Variablen werden zu einem grafischen Element zugeordent.

Typische Glyph-Visualisierungen sind z.B. Chernoff Faces, Colour Icons, Stick Figures, Star Glyphs, Pies und Metapher-Grafiken. Für eine ausführliche Zusammenfassung von Glyph-Visualisierungstechniken verweisen wir auf die Arbeit von Ward (2008). In Abbildung 4.5 werden die Daten in Beispiel 4.1.1 in Star Glyphs (Grafik oben) und Bar Glyphs (Grafik unten) dargestellt. Für die Glyph-Visualisierung hier wird die statistische Software **Gauguin** (Gribov & Unwin (2006)) verwendet, die noch zusätzliche Clusteringfunktionen für die Aufdeckung der Datenstruktur anbietet. Nähere Information dazu befindet sich unter http://rosuda.org/software/Gauguin/gauguin.html.

Glyph-Visualisierungstechniken wurden in den letzten Jahren häufig in der multivariaten Visualisierung eingesetzt. Ein Problem der Glyph-Visualisierung liegt in der Verzerrung der Wahrnehmung, was bedeutet, dass manche Variablen bzw. die Zusammenhänge von manchen Variablen einfacher wahrzunehmen als die von anderen Variablen sind, z.B., in diesem Sinne haben die benachbarten Variablen den Vorteil gegenüber den nichtbenachbarten Variablen in Bar Glyphs in Abbildung 4.5. Da die Werte der Daten in einer Glyph-Darstellung nicht direkt ablesbar sind, ist in der Praxis eine interaktive Glyph-Darstellung zu empfehlen. Die Software **Gauguin** bietet ein volles Linking zwischen den Glyphs und anderen Grafiken, damit die Benutzer die empirischen Daten möglichst vollständig untersuchen können.

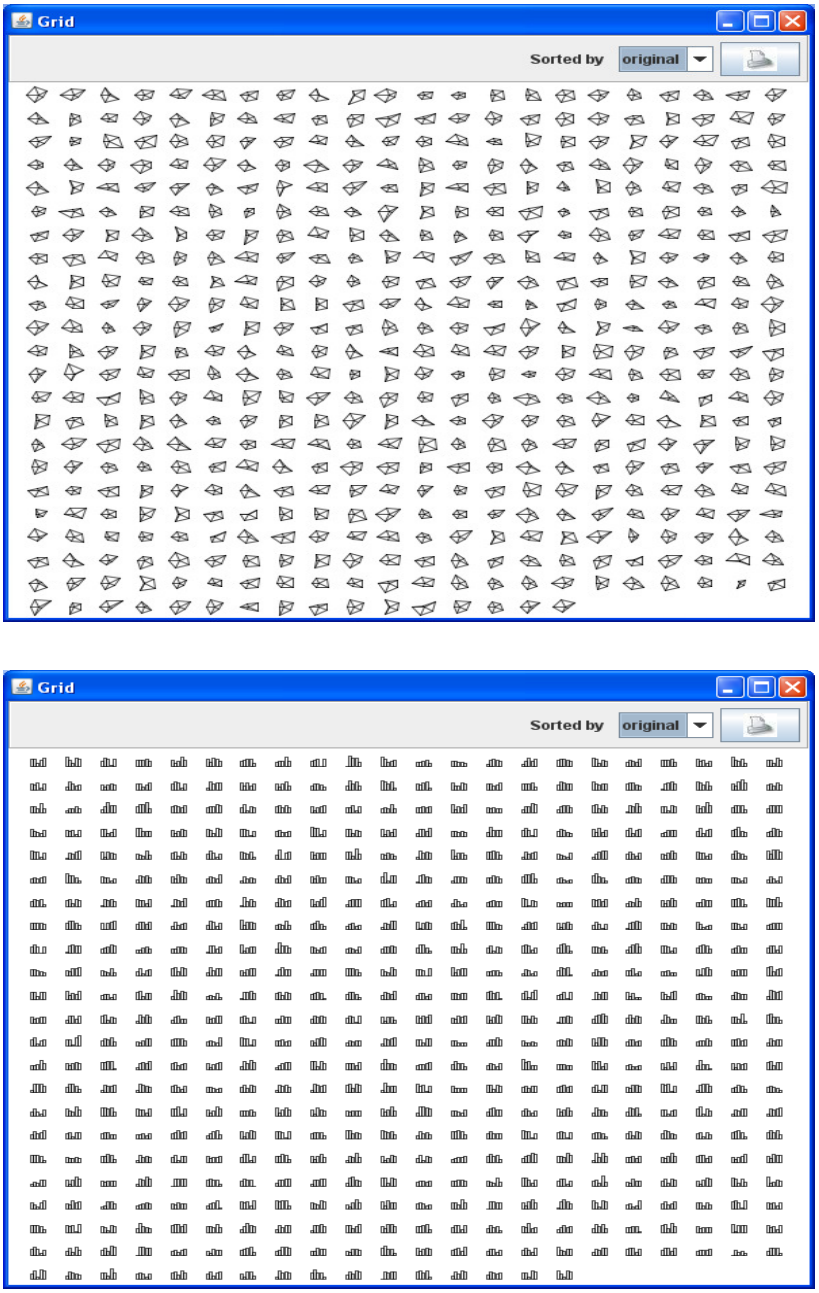

Abbildung 4.5: Daten in Beispiel 4.1.1 in Star Glyphs (Grafik oben) und Bar Glyphs (Grafik unten)

Heatmap

Der Begriff von Heatmap (auch Matrix Visualisierung genannt) wurde zuerst von Bertin (1967) in die Datenanalyse eingeführt. Gegeben seien die Daten D in Matrixform $D \in R^{n\times d}$, wobei n bzw. d für die Anzahl von Daten bzw. Variablen steht, dann wird eine Heatmap grundsätzlich wie folgt konstruiert:

- Man wählt eine geeignete Permutation P_1 für die Daten und P_2 für die Variablen;
- Man ordnet die Zeilen und Spalten von D nach P_1 und P_2 und erhält eine neue Datenmatrix D', z.B., man ordnet üblicherweise die Zeilen von D nach einem hierarchischen Dendrogramm und die Spalten von D nach Korrelationen der Variablen;
- Man stellt D' grafisch in zweidimensionalen, rechteckigen, $n \times d$ Gitterpunkten dar, wobei der Gitterpunkt in i-ter Zeile und j-ter Spalte der Größe des Eintrags $d'_{i,j}$ in D' entsprechend eingefärbt ist;
- Bemerkung: In der praktischen Anwendung sind die Zeilen und Spalten von D oft umtauschbar, z.B., bei der Anwendung der Heatmap auf den Microarray Daten.

Im Folgenden zeigt man zwei Beispiele der Datenvisualisierung in der Heatmap. Dabei wird das Rot-Weiß-Blau Farbspektrum verwendet, d.h., die kleinen (großen) Werte werden rot (blau) eingefärbt, und die Werte, die nah am Mittelwert liegen, werden in weißer Farbe dargestellt. Abbildung 4.6 zeigt die Daten in Beispiel 4.1.1 in einer Heatmap, wobei die Zeilen der Datenmatrix in der linken (rechten) Grafik nach dem Single (Complete) Linkage Dendrogramm geordnet worden sind. Man sieht in Abbildung 4.6, dass 5 große Modi der Daten in der rechten Grafik zu erkennen sind, während die linke Grafik nur 3 große Modi zeigt. Zu bemerken ist, dass das Complete Linkage Verfahren in diesem Fall in der Tat ein ziemlich gutes Resultat liefert, das aber nicht durch die Heatmap vollständig widerspiegelt worden ist. Den Vergleich des Resultats aus dem Complete Linkage Clustering mit den vorgegebenen 6 Komponenten stellt man im Fluctuationsdiagramm in Abbildung 4.7 dar.

Eine Heatmap Darstellung der Daten hängt stark von P_1 und P_2 ab, d.h., man erhält unterschiedliche Heatmaps unter verschiedenen P_1 und P_2. Abbildung 4.8 zeigt ein anderes Beispiel für die Datenvisualisierung in der Heatmap, wobei die **Italienwein** Daten in Beispiel 3.3.2 in Heatmaps gezeigt werden. Man sieht in Abbildung 4.8, dass im Unterschied zur vorgegebenen Weinklasse keine Clusterstruktur der Daten in der Heatmap zu erkennen ist. Der Grund liegt darin, dass wie es in Abschnitt 3.3 steht ein hierarchisches Clustering Verfahren mit $d(x, y) = ||x - y||_2$ die wahre Clusterstruktur der Daten nicht gut widerspiegeln kann.

In der Praxis wird diese Visualisierungstechnik sehr breit für die Aufdeckung der Clusterstruktur in den empirischen Daten angewendet. Die wichtigste Anwendung von der Heatmap-Visualisierung befindet sich wohl in der Visualisierung von den Microarray Daten. Die folgenden zwei Punkte sind wichtig bei der Heatmap-Visualisierung:

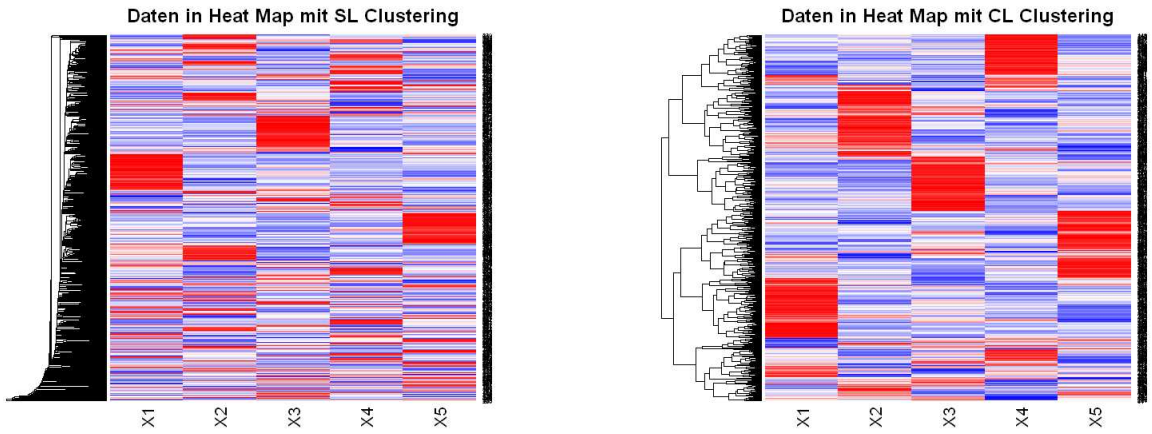

Abbildung 4.6: Daten in Beispiel 4.1.1 in Heatmap. Permutation der Zeilen der Datenmatrix nach dem Single Linkage Dendrogramm (Grafik links) und Complete Linkage Dendrogramm (Grafik rechts)

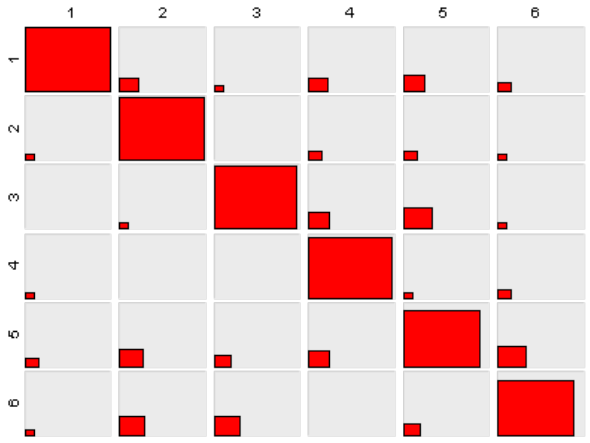

Abbildung 4.7: Vergleich des Resultats aus dem Complete Linkage Verfahren mit den vorgegebenen 6 Komponenten in Beispiel 4.1.1 im Fluctuationsdiagramm

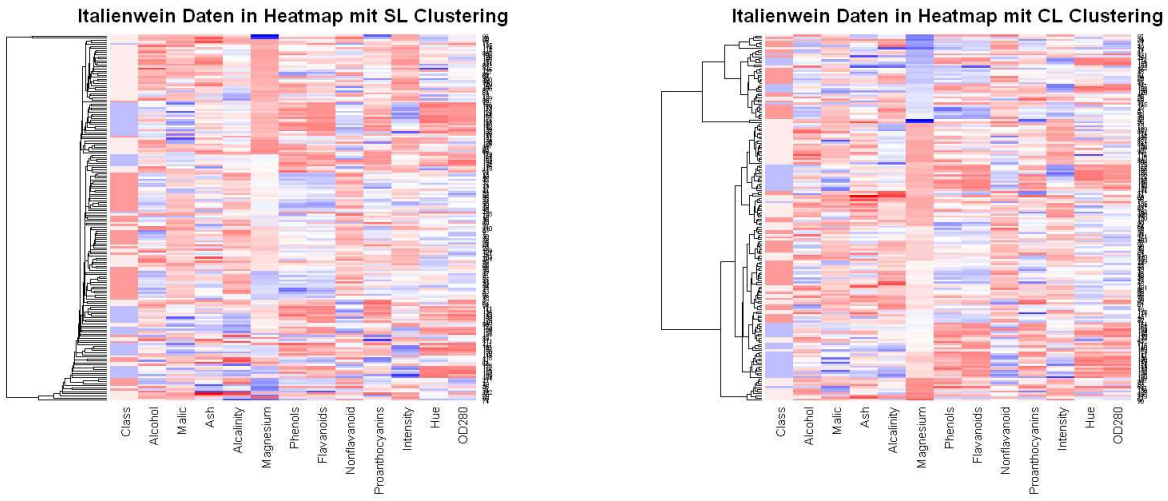

Abbildung 4.8: **Italienwein** Daten in Beispiel 3.3.2 in Heatmaps. Permutation der Zeilen der Datenmatrix nach dem Single Linkage Dendrogramm (Grafik links) und Complete Linkage Dendrogramm (Grafik rechts)

- Die Nützlichkeit einer Heatmap-Darstellung hängt stark von dem dahinter stehenden statistischen Modell ab. Es ist wichtig, ein für die Daten geeignetes statistisches Modell zu wählen;
- Die Information aus dem statistischen Modell soll durch die Permutation der Zeilen bzw. Spalten von D und durch die Einfärbung der Gitterpunkte möglichst vollständig widerspiegelt werden. Es ist wichtig, eine dafür geeignete Permutation und ein passendes Farbspektrum zu bestimmen.

Außer den obigen zwei Punkten ist eine interaktive Heatmap-Darstellung der Daten zu empfehlen. Die Software SEURAT (Gribov (2009)) aus ROSUDA der Universität Augsburg bietet z.B. eine volle interaktive Heatmap Visualisierung, wodurch man möglichst viel Information aus den Daten extrahieren kann.

Projection Pursuit

Projection Pursuit steht für eine Klasse statistischer Verfahren zur Entdeckung der unbekannten Struktur in multivariaten Daten. Der Name „Projection Pursuit“ wurde von Friedman & Tukey (1974) vergeben. Die Grundidee von Projection Pursuit liegt darin, dass man die interessanten Projektionen der multivariaten Daten durch Maximierung einer Zielfunktion (üblicherweise als Projection Pursuit Index genannt) findet und untersucht, weil die Struktur der multidimensionalen Daten in der Regel nicht direkt beschreibbar ist. Eine Projektion der Daten definiert man wie folgt: Sei X eine $n \times d$ Datenmatrix, dann bekommt man eine d'-dimensionale Projektion Y von X durch $Y = XA$, wobei A eine $d \times d'$ orthonormale Projektionsmatrix ist. Eine interessante Projektion der

Daten bezieht sich in erster Linie auf die Projektion mit einer Nicht-Normal-Struktur der projizierten Daten, weil die meisten Projektionen der multidimensionalen Daten eine Normal-Struktur zeigen. Um die interessanten Projektionen der Daten zu finden, wurde in den letzten Jahren eine Menge von Projection Pursuit Indizes vorgeschlagen. In **GGobi** werden die Indizes Holes, Central Mass, PCA, LDA, Gini-C und Entropy-C verwendet. Für eine ausführliche Beschreibung von Projection Pursuit bzw. **GGobi** verweisen wir auf die Arbeit von Friedman & Tukey (1974), Huber (1985), Friedman (1987), Jones et al. (1987), Hall (1989), Nason (1992, 1995), Cook et al. (1993, 1995, 1997, 2007, 2008) u.a. und unter http://ggobi.org/.

Im Folgenden wird dieses Verfahren anhand des Beispiels 4.1.1 kurz vorgestellt und mittels statistischer Software **GGobi** (Swayne et al. (1999)) veranschaulicht. Hier versucht man, die Clusterstruktur der Daten in Beispiel 4.1.1 zu untersuchen, indem man die interessanten uni- und bivariaten Projektionen der Daten anhand des Holes Indexes untersucht. Ein Holes Index wird wie folgt definiert:

$$I_{Holes}(A) = \frac{1 - \frac{1}{n}\sum_{i=1}^{n} exp(-\frac{1}{2}y_i'y_i)}{1 - exp(-\frac{d'}{2})} \tag{4.3}$$

wobei $[y_1, y_2, ..., y_n]^T = Y$ die $n \times d'$ Matrix der Projektion der Daten ist. Im Vergleich zu den anderen Zielfunktionen findet man durch Maximierung von $I_{Holes}(A)$ die Projektionen mit einem „Gap“ zwischen zwei Clustern der projizierten Daten (Cook et al. (2008)). Abbildung 4.9 zeigt beispielhaft zwei interessante 1D Projektionen der Daten in Beispiel 4.1.1 in Bezug auf den Holes Index. Ein paar Erklärungen dazu:

- Daten aus verschiedenen Komponenten werden unterschiedlich eingefärbt (vgl. Grafik oben rechts). Abbildung 4.10 zeigt die eingefärbten Daten in der Scatterplot Matrix in **GGobi**;
- Zwei eindimensionale Projektionen der Daten mit Projektionsmatrizen $A_1 = (-0,628; 0,238; 0,273; -0,609; 0,322)^T$ (Grafik mittel rechts) und $A_2 = (0,085; -0,664; 0,143; -0,721; 0,105)^T$ (Grafik unten rechts) werden gezeigt;
- $I_{Holes}(A_1) = 0,814$ (vgl. Grafik mittel links) und $I_{Holes}(A_2) = 0,818$ (vgl. Grafik mittel rechts);
- In der Grafik mittel rechts befindet sich ein „Gap“ in der geschätzten Dichtefunktion der 1D projizierten Daten, mit dem die gelbe Gruppe in zwei Teile gesplittet worden ist;
- In der Grafik unten rechts befindet sich ein „Gap“ in der geschätzten Dichtefunktion der 1D projizierten Daten, mit dem die gelbe Gruppe eigentlich von den anderen Gruppen getrennt worden ist.

Ein paar Bemerkungen dazu:

- Die Existenz von dem „Gap“ hängt von dem Glättungsparameter ab (vgl. Grafik oben links in Abbildung 4.9). In der Regel ist es nötig zu überprüfen, ob der „Gap“ signifikant ist;
- Die gelbe Gruppe hat einen großen Einfluss auf das Ergebnis (vgl. auch Abbildung 4.10), weil die Gruppe von den anderen 5 Gruppen umringt ist;
- Für die Entdeckung der Clusterstruktur in den Daten in Beispiel 4.1.1 ist es nötig, die zwei in der Grafik unten rechts in Abbildung 4.9 gezeigten getrennten Teilmengen der Daten weiter zu untersuchen;
- Falls keine a priori Information über die Gruppierung der Daten vorhanden ist, ist die Clusterstruktur in den Daten rein durch Untersuchung der interessanten 1D Projektionen der Daten nicht aufzudecken. Dafür besteht die Notwendigkeit, die Information aus einem Clusteringverfahren mit in die Visualisierung einzubeziehen.

Abbildung 4.11 zeigt zwei interessante 2D Projektionen der Daten in Beispiel 4.1.1 in Bezug auf den Holes Index. In Abbildung 4.11 sieht man einerseits, dass die ersten vier Gruppen (von links nach rechts, vgl. Grafik oben rechts) eigentlich in beiden interessanten 2D Projektionen gut getrennt worden und wegen der Existenz der Daten aus den anderen zwei Gruppen (gelb und grau eingefärbt) nicht erkennbar sind, andererseits dass die Clusterstruktur der Daten in den beiden 2D Projektionen ohne a priori Information nicht direkt aufzudecken ist. Ein sinnvoller Vorschlag für das unsupervised Clustering mittels Projection Pursuit liegt in der wiederholten Bi-partitionierung der projizierten Daten jeweils in der interessantesten Projektion der multidimensionalen Daten (vgl. auch Miasnikov et al. (2004)).

Oben wurden ein paar Visualisierungsmethoden für die multivariaten Daten anhand von Beispielen vorgestellt. In der Tat ist es üblich und notwendig, eine Visualisierungstechnik zusammen mit einem geeigneten statistischen Modell zu benutzen, so dass man die Datenstruktur durch die Visualisierung der Daten bzw. der Information aus dem statistischen Modell aufdecken kann. Im nächsten Abschnitt wird anhand der Beispiele 4.1.1, 3.3.1 und 3.3.2 gezeigt, dass man die unbekannte Datenstruktur durch grafische Darstellung der Information aus Dichteschätzung gut untersuchen kann. In Abschnitt 4.3 wird gezeigt, dass die Clusterstruktur in den Daten in Beispiel 4.1.1 durch geeignete Visualisierung der Information aus dem Dichteschätzer basierten hierarchischen Clustering aufzudecken ist.

4.2 Dichteschätzung und Visualisierung

In diesem Abschnitt wird die Nutzung der Information aus Dichteschätzung in der Datenvisualisierung diskutiert. In den folgenden Situationen ist es hilfreich, die Information aus Dichteschätzung mit in der Datenvisualisierung einzubeziehen:

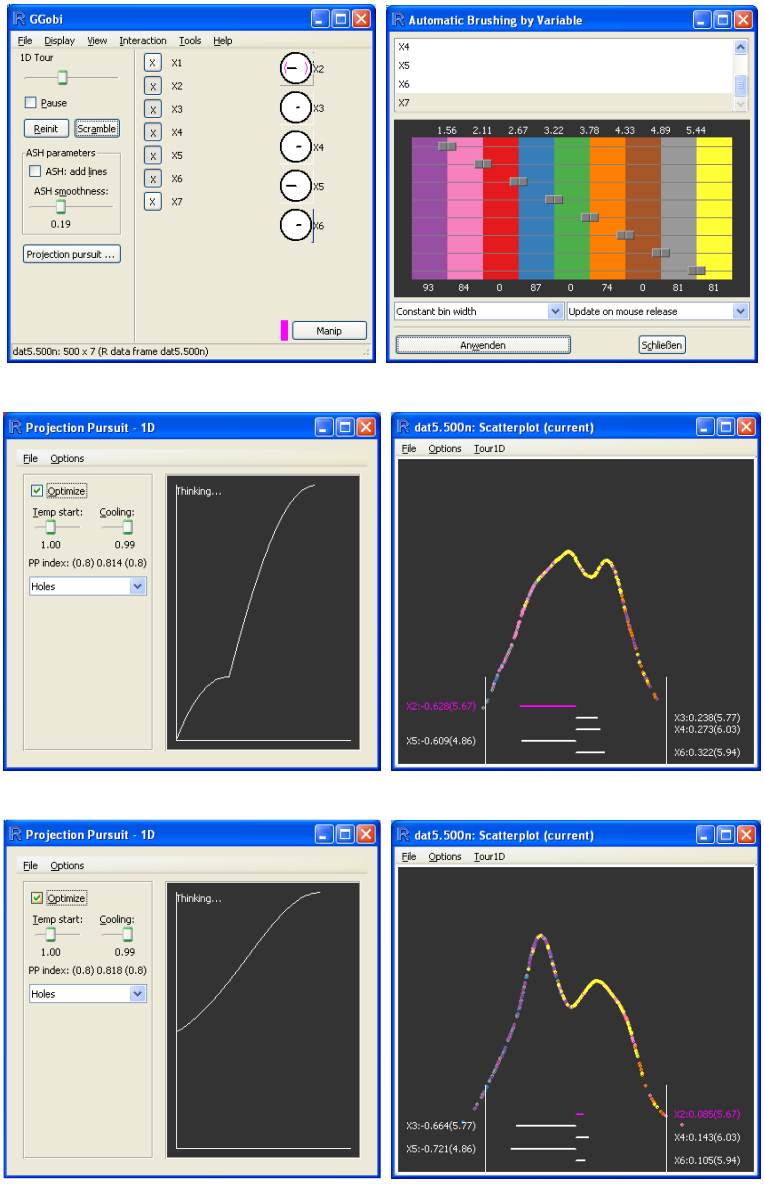

Abbildung 4.9: Zwei interessante 1D Projektionen der Daten in Beispiel 4.1.1 in Bezug auf den Holes Index in **GGobi**

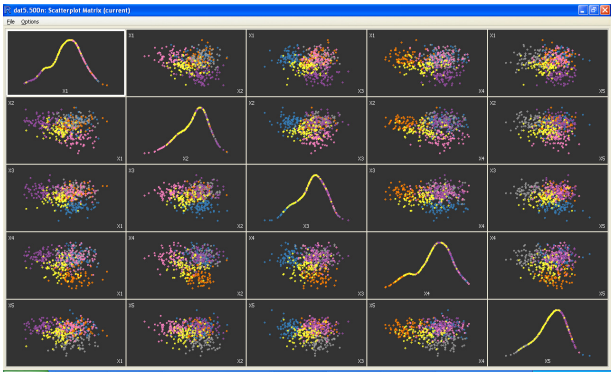

Abbildung 4.10: Eingefärbte Daten in Beispiel 4.1.1 in der Scatterplot Matrix in **GGobi**

- Die Daten sind grafisch nicht direkt darstellbar, z.B., multivariate Daten mit Dimension größer 3;
- Das Muster der Daten ist in einer punkt-orientierten grafischen Darstellung schwierig erkennbar, z.B., im Fall eines großen bivariaten Datensatzes;
- Man wollte die Verteilung der Daten besser zeigen, z.B., durch ein Histogramm in der univariaten Datenanalyse.

Das 1D Histogramm ist wohl eine der populärsten grafischen Methoden für die Darstellung der Daten aus einer 1D stetigen Variable. Im 2-3D Fall sind die punkt-orientierten grafischen Methoden nicht gut anwendbar, wenn der Umfang der Daten groß ist, weil in diese Situation die Datenstruktur wegen der Überfüllung von Pixeln in der Grafik kaum erkennbar ist. Aus diesem Grund kann man die Information aus einem geeigneten statistischen Modell auf Basis der zu untersuchenden Daten grafisch darstellen anstatt der originalen Daten, z.B., einen geeigneten Dichteschätzer in Contour Linien oder im Imageplot. In diesem Abschnitt wird anhand der Beispiele 4.1.1, 3.3.1 und 3.3.2 diskutiert, wie man die Information aus Dichteschätzung und darauf basiertem Clusteringverfahren für die multivariate (Dimension der Daten größer 3) Datenanalyse grafisch darstellt. Das Ziel des Einbezugs der multivariaten Dichteschätzung in die Visualisierung liegt hauptsächlich darin, die Cluster- bzw. Modal-Struktur der Daten zu untersuchen.

Dichteschätzer mit dem Parallel Koordinaten Plot

Für den Einbezug der Information aus Dichteschätzung in den Parallel Koordinaten Plot wird hier Folgendes vorgeschlagen:

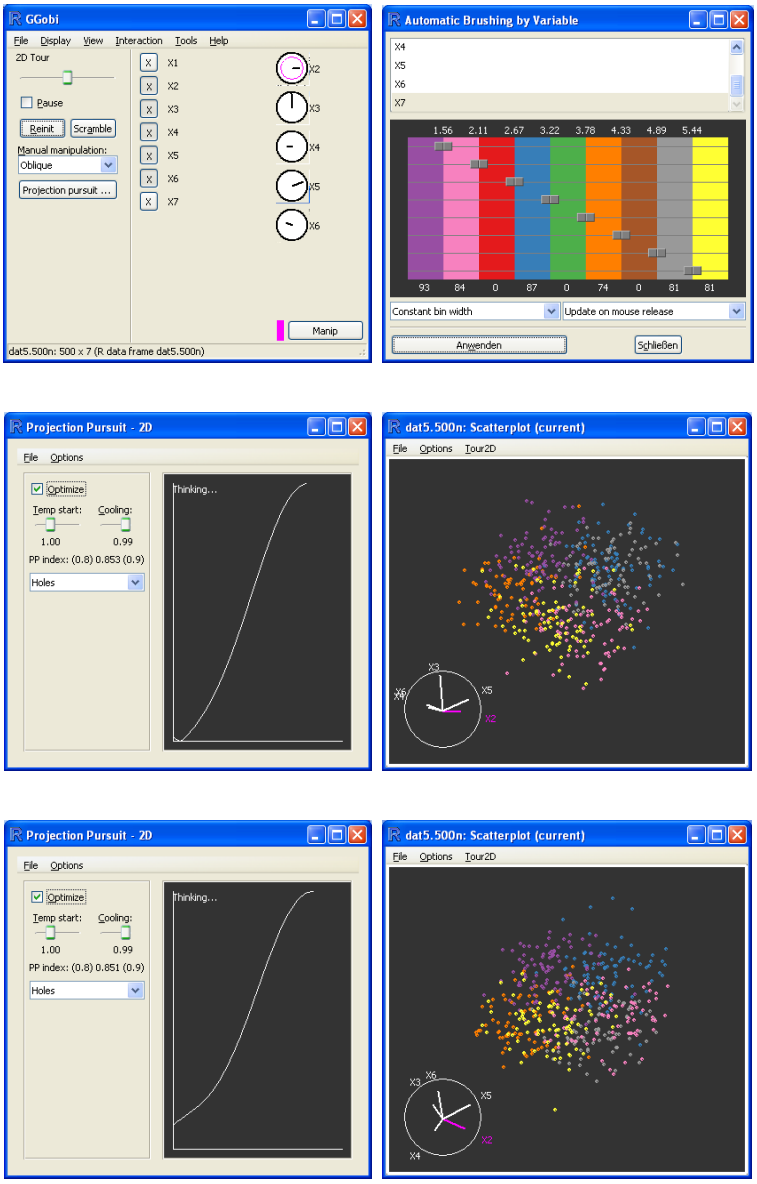

Abbildung 4.11: Zwei interessante 2D Projektionen der Daten in Beispiel 4.1.1 in Bezug auf den Holes Index in **GGobi**

- Man wählt ein geeignetes Farbspektrum für das Zeichnen der Linien im Parallel Koordinaten Plot aus;
- Die Linien mit $\hat{f}(x) < \lambda$ werden in der Grafik unterdrückt, d.h., weiß oder leicht eingefärbt, wobei $\hat{f}(x)$ die geschätzte Dichte ist;
- Die Linien mit $\hat{f}(x) \geq \lambda$ werden der Größe von $\hat{f}(x)$ nach eingefärbt.

Abbildung 4.12 zeigt die Daten in Beispiel 4.1.1 im Parallel Koordinaten Plot, wobei

- das beste Modell mit 6 Komponenten aus dem Model Based Clustering von Fraley & Raftery (2002) als geschätzte Dichte $\hat{f}$ verwendet wird;
- das Farbspektrum (weiß - blau) benutzt wird, was bedeutet, dass die Linien mit kleinen (großen) Werten von $\hat{f}$ leicht (blau) eingefärbt werden;
- ein gewisser Teil der Daten mit kleinen Werten von $\hat{f}$ unterdrückt (weiß eingefärbt) wird (10% in der Grafik oben links, 20% oben rechts, 30% unten links 40% unten rechts);
- die Linien der Größe von $\hat{f}$ nach eingefärbt werden.

In Abbildung 4.12 ist die Datenstruktur klarer zu erkennen im Vergleich zu Abbildung 4.3. In der Tat wird die wahre Dichte durch das gemischte Modell aus dem Model Based Clustering gut widerspiegelt, wenn man Abbildung 4.12 mit Abbildung 4.13, in der die Linien im Parallel Koordinaten Plot nach der Größe der wahren Dichte eingefärbt sind, vergleicht. Zu bemerken ist, dass der Einbezug der Information aus Dichteschätzung in den Parallel Koordinaten Plot nur die Möglichkeit bietet, die Daten aus einem anderen Sichtwinkel zu betrachten. Um die Datenstruktur weiter zu untersuchen, ist eine interaktive Darstellung der Daten im Parallel Koordinaten Plots nötig. Dafür stehen die Softwares **Manet**, **Mondrian**, **iplots** Paket in **R** aus ROSUDA der Universität Augsburg zur Verfügung. Weitere Information befindet sich unter http://rosuda.org/software/.

Dichteschätzung mit dem Einbettungsverfahren

Multivariate Daten können durch das Einbettungsverfahren auf Basis eines geeigneten Dichteschätzers visualisiert werden. Im Folgenden wird ein Einbettungsverfahren vorgestellt, in dem die „Querschnitte“ des Dichteschätzers gezeigt werden. Dieses Einbettungsverfahren ist für 3 bis 6 dimensionale Daten gut geeignet und bietet die Möglichkeit an, dass die Benutzer im multivariaten Raum „surfen“ können, um die interessanten Regionen zu finden und näher zu untersuchen. Dabei geht man wie folgt vor:

- Man teilt die Variablen (nach Korrelationsverhältnissen, durch die Faktoranalyse oder per Zufall) in zwei Subgruppen V_1 und V_2 auf;
- Dementsprechend wird der ganze Datenraum in zwei Subräumen R_1 und R_2 aufgeteilt;

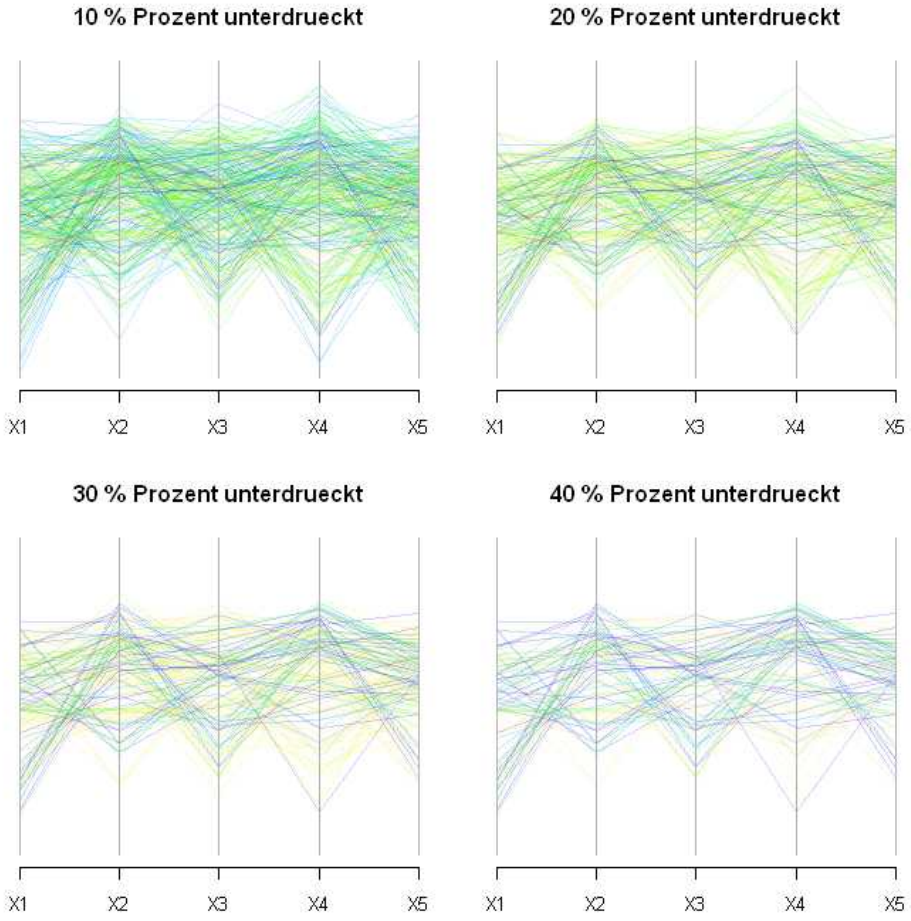

Abbildung 4.12: Daten in Beispiel 4.1.1 im Parallel Koordinaten Plot. Die Linien mit kleinen (großen) Werten von $\hat{f}_m$ werden leicht (blau) eingefärbt. Daten unterdrücken (weiß eingefärbt): 10% in der Grafik oben links, 20% oben rechts, 30% unten links und 40% unten rechts.

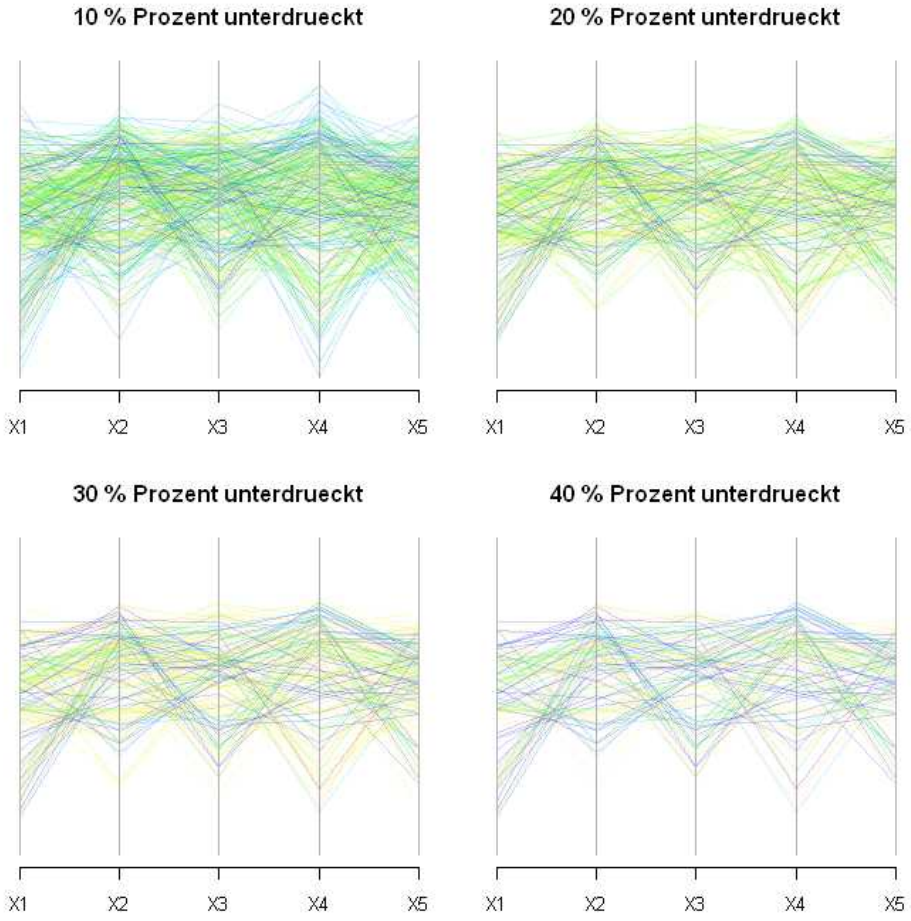

Abbildung 4.13: Daten in Beispiel 4.1.1 im Parallel Koordinaten Plot. Die Linien mit kleinen (großen) Werten von der wahren Dichte werden leicht (blau) eingefärbt. Daten unterdrücken (weiß eingefärbt): 10% in der Grafik oben links, 20% oben rechts, 30% unten links und 40% unten rechts.

- Man bestimmt die Gitterpunkte $G_1 \subset R_1$;
- Man bettet R_1 in R_2 ein, d.h., jeder Punkt in R_2 steht dann für eine 2- oder 3 dimensionale Hyperebene in R^d;
- Man wählt eine Route R_o (eine stetige Kurve) im R_2 als einen Slider aus;
- Man berechnet die Dichteschätzer $\hat{f}((g,r)^T)$, wobei $g \in G_1$, $r \in r_o$ mit $r_o \subset R_o$;
- Wenn man sich entlang R_o „bewegt“, erhält man eigentlich eine Serie von „Querschnitten“ des Dichteschätzers;
- Man findet die interessanten Regionen und untersucht sie näher.

Die Grundidee für das Einbettungsverfahren liegt darin, dass man die wahre Clusterstruktur der Daten durch Betrachtung der Querschnitte von $\hat{f}$ erkennen kann, weil multivariate Daten im Datenraum in der Regel spärlich verteilt sind. Es ist auch zu erwarten, dass diese Erkennung nicht stark von $\hat{f}$ beeinflusst wird. Im Folgenden wird diese Einbettungsmethode anhand des Beispiels 4.1.1 veranschaulicht. Als Beispiel wird das Einbettungsverfahren wie folgt konstruiert:

- Man verwendet hier zwei Dichteschätzer als Basis für das Einbettungsverfahren, die sind $\hat{f}_m((g,r)^T)$, das beste Modell aus dem Model Based Clustering, und $\hat{f}_k((g,r)^T)$, der Kerndichteschätzer mit Normal-Reference Bandbreiten;
- Man teilt die fünf Variablen der Daten per Zufall in zwei Subgruppen mit $V_1 = (Y_3, Y_4)$ und $V_2 = (Y_1, Y_2, Y_5)$;
- Man wählt 41×41 äquidistante Gitterpunkte in R_1. Die linke Grafik in Abbildung 4.14 zeigt G_1;
- Als r_o werden 16 äquidistante Punkte entlang der Diagonal von R_2 ausgewählt. Die rechte Grafik in Abbildung 4.14 zeigt r_o bzw. R_o;
- Abbildung 4.15 zeigt die Querschnitte von $\hat{f}_m((g,r)^T)$ im Imageplot;
- Abbildung 4.16 zeigt die Querschnitte von $\hat{f}_k((g,r)^T)$ im Imageplot;

Zum Vergleichen stellt man die Querschnitte der wahren Dichte $f((g,r)^T)$ im Imageplot in Abbildung 4.17 dar. Man sieht in Abbildung 4.15-4.17 folgendes:

- Die Modalstruktur wird in Abbildung 4.15 richtig identifiziert;
- Die Modalstruktur wird in Abbildung 4.16 auch richtig (außer am Rand) identifiziert, obwohl $\hat{f}_k$ einen schlechten Schätzer für f darstellt.

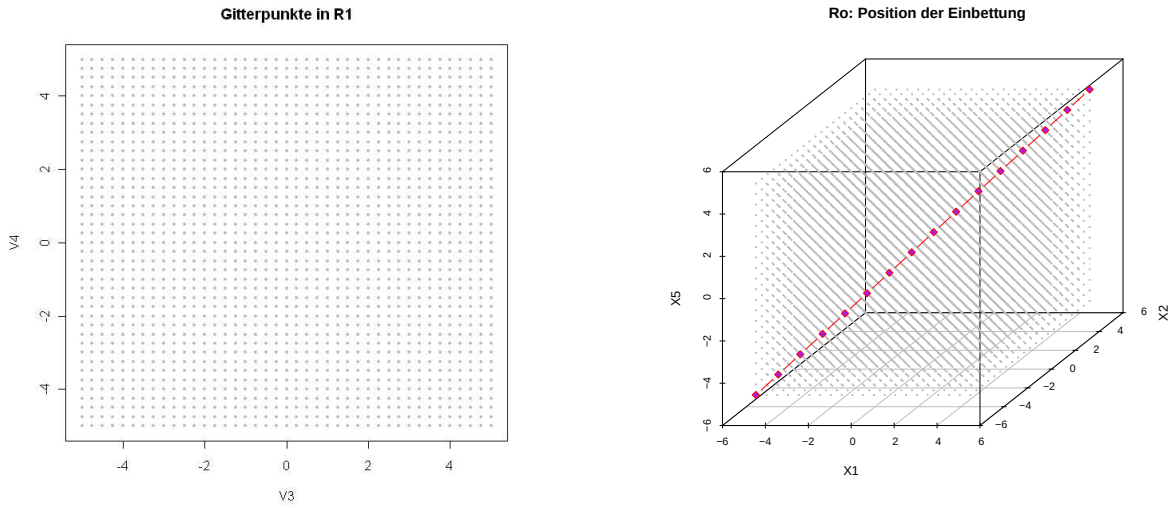

Abbildung 4.14: 41 × 41 äquidistante Gitterpunkte in R_1 (Grafik links) und 16 äquidistante Punkte entlang der Diagonal von R_2 als r_o (Grafik rechts)

In diesem Fall kann man die interessanten Regionen dadurch näher untersuchen, indem man $\hat{f}_m$ bzw. $\hat{f}_k$ an mehreren Punkten in diesen Regionen oder an anderen Punkten aus Routen in der Nähe von diesen Regionen berechnet. Beispielsweise kann man 16 äquidistante Gitterpunkte aus R_o in der Strecke von $0,33 \leq Y_1 = Y_2 = Y_5 \leq 2,33$ als r'_o auswählen und die entsprechenden Querschnitte von $\hat{f}_m$, $\hat{f}_k$ und f in Abbildung 4.18, 4.19 und 4.20 darstellen. In Abbildung 4.15-4.20 sieht man, dass zwei Modi $(1,5; 1,5; -1,5; 1,5; 1,5)^T$ und $(1,5; 1,5; 1,5; -1,5; 1,5)^T$ durch dieses Einbettungsverfahren auf Basis der beiden Dichteschätzer richtig identifiziert worden sind. Abbildung 4.21 zeigt das Bestimmtheitsmaß aus zwei einfachen linearen Modellen $f((g,r)^T) \sim \hat{f}_m((g,r)^T)$ und $f((g,r)^T) \sim \hat{f}_k((g,r)^T)$ für $g \in G_1$ jeweils an r_o (Grafik links) r'_o (Grafik rechts). In Abbildung 4.21 ist zu erkennen, dass f an diesen Stellen durch $\hat{f}_m$ gut geschätzt und im Zentralbereich auch durch $\hat{f}_k$ widerspiegelt wird.

Oben wurde gezeigt, dass zwei Modi der Daten in Beispiel 4.1.1 durch das Einbettungsverfahren richtig identifiziert worden sind. Um alle möglichen Modi der Daten zu entdecken, sind mehrere „Querschnitte“ eines Dichteschätzers zu untersuchen. Die Schwierigkeit dieses Einbettungsverfahrens liegt darin, dass man eine unendliche Zahl möglicher „Querschnitte“ hat. Dieses Problem kann dadurch gelöst werden, indem man die 2 oder 3D Hyperebenen, auf denen die geschätzten Dichten an allen Gitterpunkten kleiner λ (ein vorgegebenes Dichteniveau) sind, ignoriert. Auf diese Art kann man sich nur auf die Regionen konzentrieren, wo sich die High Density Cluster befinden. Eine interaktive Darstellung ist beim Anwenden dieses Einbettungsverfahrens stark zu empfehlen, damit man mit R_1 in R_2 „surfen“ kann, um die unbekannte Datenstruktur vollständig zu untersuchen.

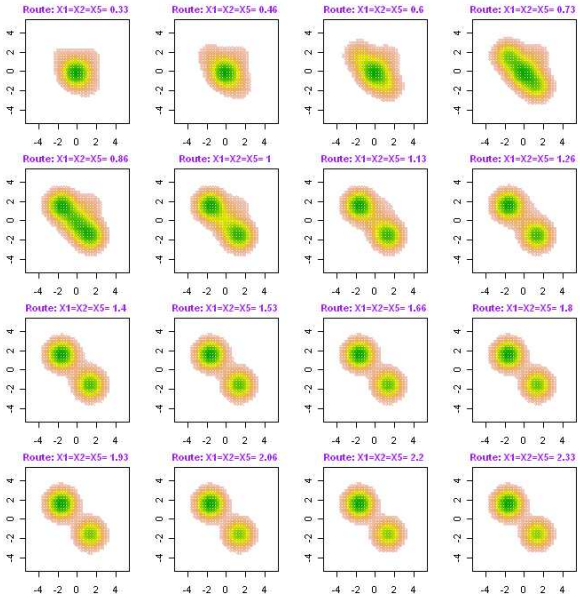

Abbildung 4.15: 16 Querschnitte von $\hat{f}_m((g,r)^T)$ im Imageplot

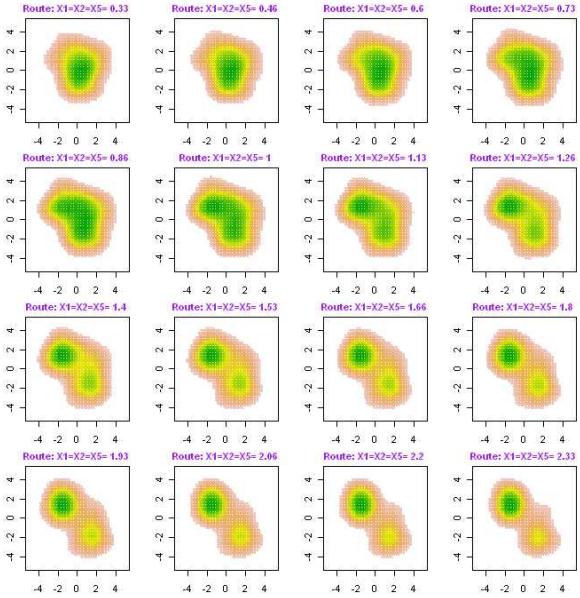

Abbildung 4.16: 16 Querschnitte von $\hat{f}_k((g,r)^T)$ im Imageplot

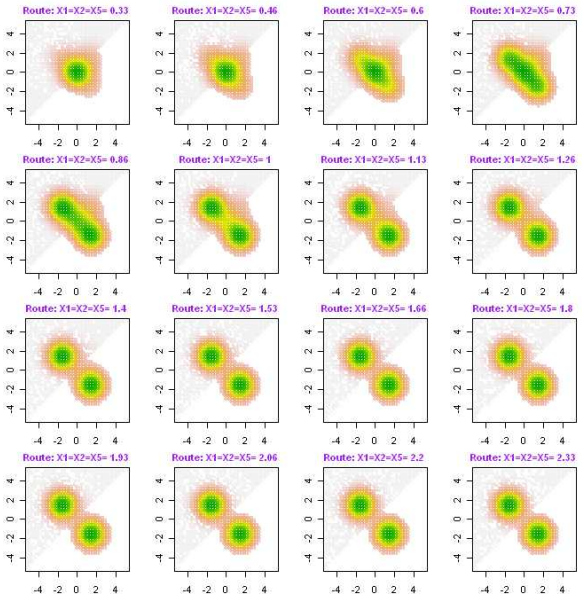

Abbildung 4.17: 16 Querschnitte von $f((g,r)^T)$ im Imageplot

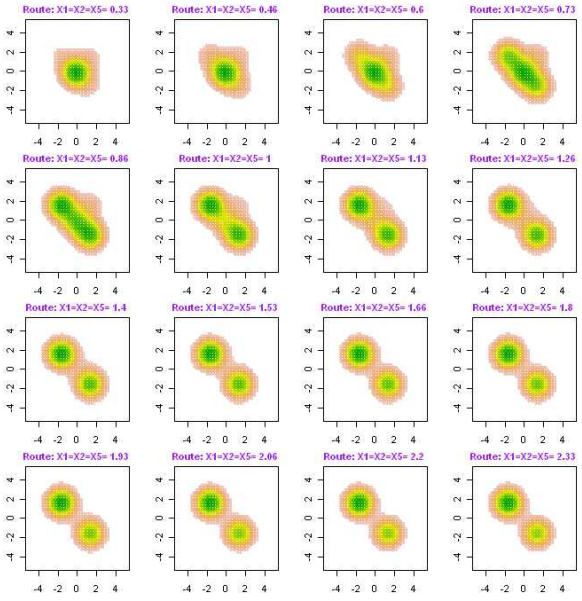

Abbildung 4.18: Querschnitte von $\hat{f}_m((g,r)^T)$ an weiteren 16 Punkten aus R_o im Imageplot

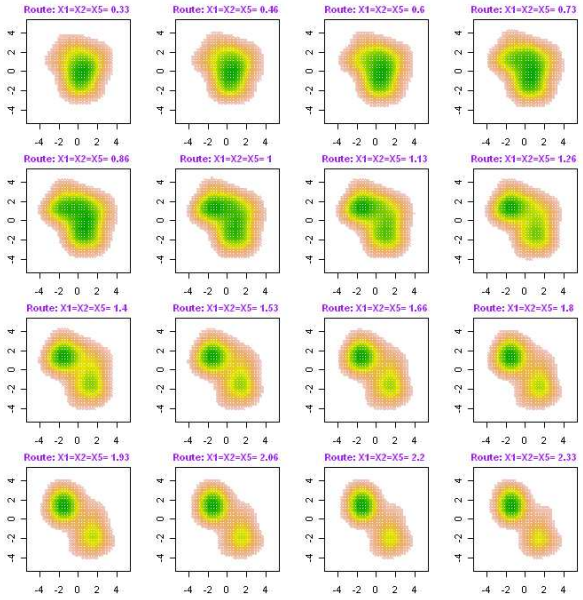

Abbildung 4.19: Querschnitte von $\hat{f}_k((g,r)^T)$ an weiteren 16 Punkten aus R_o im Imageplot

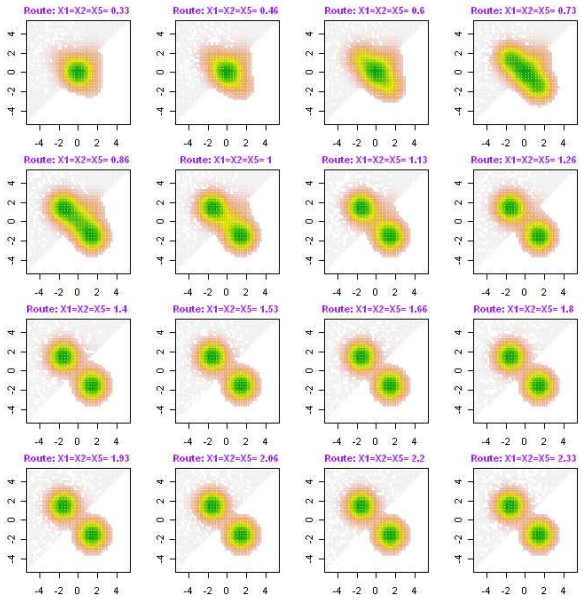

Abbildung 4.20: Querschnitte von $f((g,r)^T)$ an weiteren 16 Punkten aus R_o im Imageplot

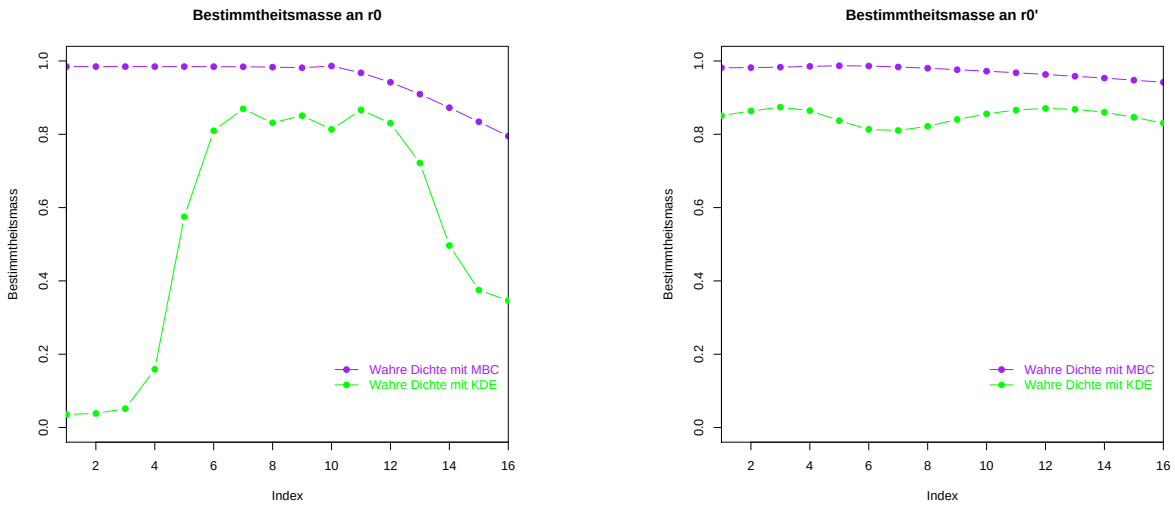

Abbildung 4.21: Bestimmtheitsmaß aus zwei einfachen linearen Modellen von $f((g,r)^T) \sim \hat{f}_m((g,r)^T)$ und $f((g,r)^T) \sim \hat{f}_k((g,r)^T)$ an r_o (Grafik links) und r'_o (Grafik rechts)

Dichteschätzung mit Heatmap

Die Heatmap-Visualisierung dient dazu, das Muster der Daten durch Permutation der Zeilen bzw. Spalten der Datenmatrix zu erkennen. Wie in Abschnitt 4.1 erwähnt spielt bei der Heatmap-Visualisierung das dahinter stehende statistische Modell eine entscheidende Rolle. Die Information aus Dichteschätzung kann dadurch mit in die Heatmap-Visualisierung einbezogen werden, indem man die Zeilen bzw. Spalten der Datenmatrix dem Dichteschätzer basierten hierarchischen Dendrogramm entsprechend permutiert. Abbildung 4.22 zeigt die Daten aus Beispiel 3.3.2 in der Heatmap, wobei die Zeilen der Datenmatrix nach dem Dichteschätzer Basierten Dendrogramm aus 2C13 (vgl. Abs. 3.3) permutiert worden sind. Man sieht in Abbildung 4.22, dass man eine gute Klassifizierung der Daten sogar mit einer Subgruppe der Variablen erhalten kann, z.B., man kann die Daten anhand der Variable *OD280* oder *Hue* in zwei Gruppen (rot und nicht rot) und dann die nicht-rote Gruppe anhand der Variable *Proline* weiter in zwei Gruppen (blau und rot) aufteilen.

In Beispiel 3.3.2 wird die vorgegebene Weinklasse durch das Resultat aus 2C13 gut widerspiegelt. Da in der explorativen Datenanalyse diese a priori Information in der Regel nicht vorhanden ist, besteht der Hauptziel der Heatmap-Visualisierung in der Exploration der Gruppierungsmöglichkeit der Daten. In diesem Sinne ist zu empfehlen, dass man die Zeilen bzw. Spalten der Datenmatrix nach unterschiedlichen statistischen Modellen permutiert und die daraus resultierten Datenmatrizen in der Heatmap darstellt, um möglichst viel Information über die Clusterstruktur der Daten zu erhalten. Im

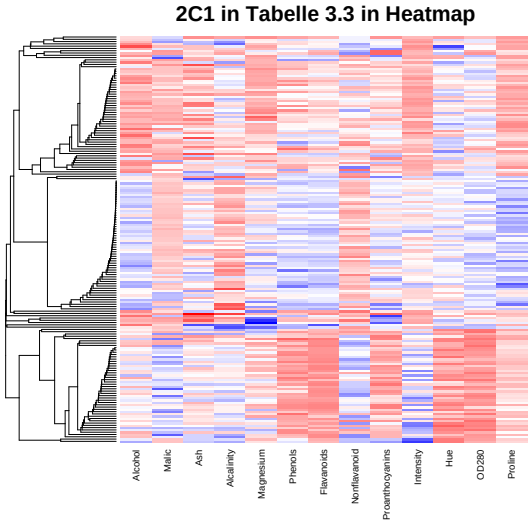

Abbildung 4.22: **Italienwein** Daten in der Heatmap. Permutation der Zeilen der Datenmatrix nach dem Dendrogramm aus 2C13 (vgl. Abs. 3.3)

nächsten Abschnitt werden zwei grafische Methoden vorgeschlagen, mit denen man die Information aus einem Dichteschätzer basierten hierarchischen Verfahren in die Datenvisualisierung einbeziehen kann.

4.3 Dichteschätzer basiertes hierarchisches Verfahren und Datenvisualisierung

In diesem Abschnitt werden zwei grafische Methoden vorgeschlagen, um die Information aus einem Dichteschätzer Basierten Cluster Dendrogramm in die Datenvisualisierung mit einzubeziehen. Die sind

1. Visualisierungsbaum: Ein Dichteschätzer Basiertes Cluster Dendrogramm kann man als einen Binärbaum betrachten (vgl. Abs. 3.3). Man verwendet hier den Algorithmus von Reingold und Tilford (1981), um diesen Binärbaum direkt zu zeichnen;

2. Methode aufgrund der Idee, die man für die Veranschaulichung von $\hat{E}_h(N)$ in Abbildung 3.27 (vgl. Abs. 3.3) verwendet hat. In diesem Abschnitt wird diese Methode als **Methode2** genannt.

Im Folgenden werden die obigen zwei Methoden anhand von Beispielen vorgestellt.

Visualisierungsbaum

Der Visualisierungsbaum ist ein nützliches Werkzeug in der multivariaten Datenanalyse. Zwei wichtigste Anwendungen des Visualisierungsbaums sind

1. CART von Breiman et al. (1984) zu zeigen (z.B. KLIMT von Urbanek (2003));
2. Resultat des Clusterings darzustellen (z.B. Dendrogramm).

Im Folgenden wird anhand des Modells aus 1S29 (vgl. Tabelle 3.2 in Abs. 3.3) gezeigt, wie man den Binärbaum aus einem Dichteschätzer basierten hierarchischen Verfahren grafisch darstellt und dadurch zum Zweck der Datenanalyse weiter benutzt. Abbildung 4.23

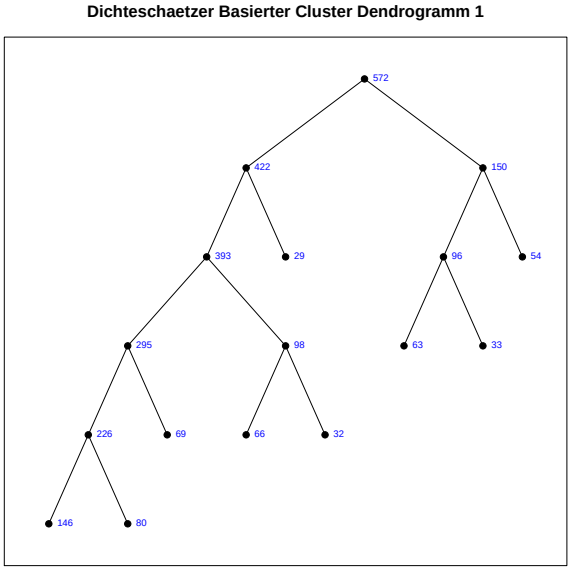

Abbildung 4.23: Grafische Darstellung des Binärbaums aus 1S29 (vgl. Tabelle 3.2 Abs. 3.3) mit dem Algorithmus von Reingold und Tilford (1981)

zeigt den Binärbaum aus 1S29 nach Reingold und Tilford (1981). Ein paar Erklärungen dazu:

- die y-Koordinate eines Knotens entspricht deren Tiefe;
- es gibt einen horizontalen Minimalabstand zwischen den Knoten auf der gleichen Tiefe;
- ein Vaterknoten liegt zentriert über seinen beiden Kindern;
- der linke (rechte) Subknoten steht für die größere (kleinere) Subgruppe der Daten;
- die Kanten des Binärbaums kreuzen sich nicht;
- die Ziffer neben des Knotens steht für die Anzahl der Daten im Cluster.

Den Visualisierungsbaum in Abbildung 4.23 kann man modifizieren, indem man Polygone anstatt von Punkten für die Darstellung der Knoten des Binärbaums verwendet. Die Fläche des Polygons ist proportional zu der Größe der entsprechenden Subgruppe der Daten. Abbildung 4.24 zeigt diesen modifizierten Visualisierungsbaum. Eine andere

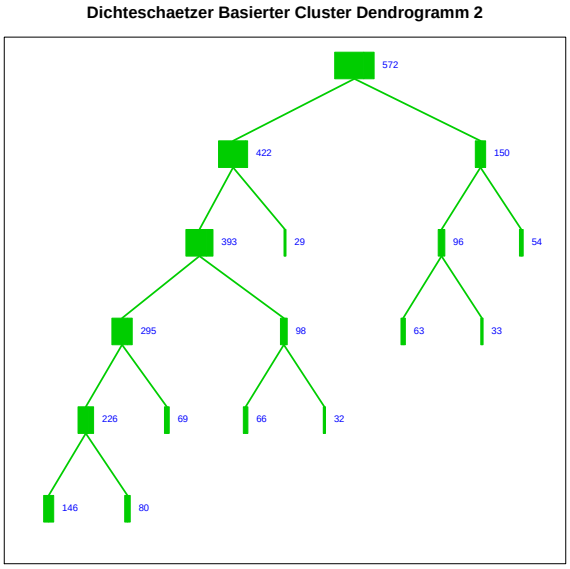

Abbildung 4.24: Modifizierter Binärbaum aus 1S29 (vgl. Tabelle 3.2 Abs. 3.3). Fläche des Polygons proportional zu der Größe der entsprechenden Subgruppe der Daten.

modifizierte Version des Visualisierungsbaums in Abbildung 4.23 stellt man in Abbildung 4.25 dar, wobei die Dicke der Kante von dem Vaterknoten zu einem seinen Kindern proportional zu der Größe der durch diesen Kinderknoten dargestellten Subgruppe der Daten ist.

Genauso wie bei anderen multivariaten Visualisierungsmethoden ist in der Praxis eine interaktive Darstellung des Visualisierungsbaums sehr hilfreich für die Untersuchung der Datenstruktur. Eine mögliche interaktive Zusammenarbeit von dem Visualisierungsbaum mit der Scatterplot Matrix und dem Parallel Koordinaten Plot durch Linking und Highlighting sieht beispielsweise wie folgt aus: Man markiert einen Knoten im Visualisierungsbaum und highlightet diese Untermenge der Daten in der Scatterplot Matrix und im Parallel Koordinaten Plot, um weitere Information über diese Untermenge der Daten zu erhalten. Zur Veranschaulichung dieses Vorgehens markiert man den rechten Knoten im Visualisierungsbaum in Abbildung 4.26 und highlightet diese Untermenge mit 54 Daten in der Scatterplot Matrix in Abbildung 4.27 und im Parallel Koordinaten Plot in Abbildung 4.28. Durch das Highlighting sieht man in Abbildung 4.27 und 4.28, dass die Daten aus diesem Cluster durch Variablen *S3* und *S4* ziemlich gut von anderen

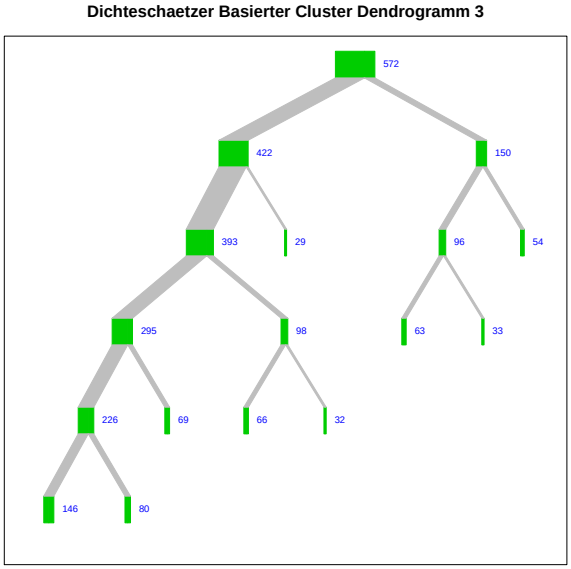

Abbildung 4.25: Modifizierter Binärbaum (2) aus 1S29 (vgl. Tabelle 3.2 Abs. 3.3). Fläche des Polygons proportional zu der Größe der entsprechenden Subgruppe der Daten. Dicke der Kante von dem Vaterknoten zu einem seinen Kindern proportional zu der Größe der durch diesen Kinderknoten dargestellten Subgruppe der Daten.

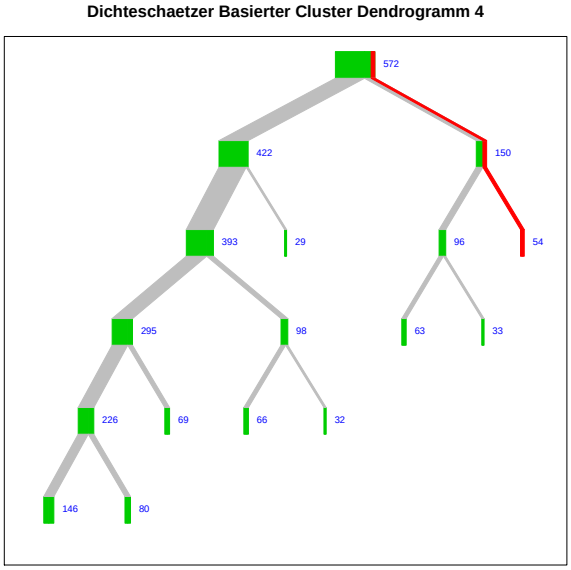

Abbildung 4.26: Markierung des rechten Knotens im Visualisierungsbaum

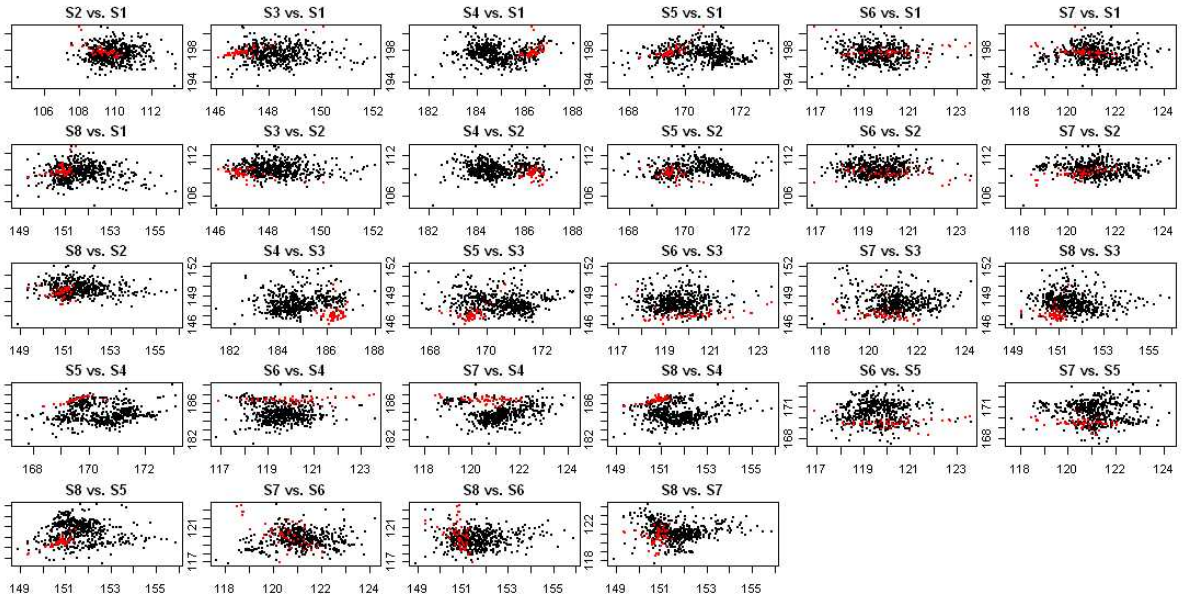

Abbildung 4.27: Highlightete Untermenge mit 54 Daten in der Scatterplot Matrix

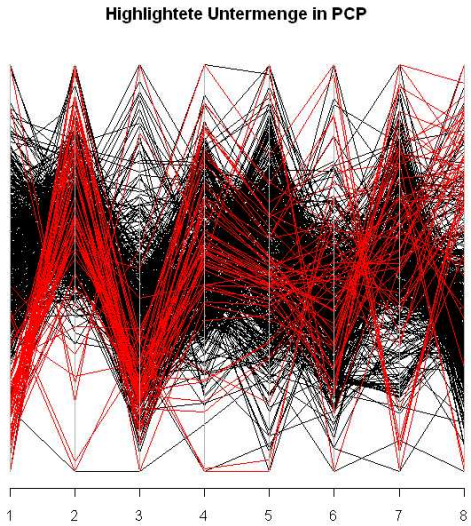

Abbildung 4.28: Highlightete Untermenge mit 54 Daten im Parallel Koordinaten Plot

Daten getrennt sind.

Der Visualisierungsbaum kann auch erweitert (reduziert) werden, indem man ein passendes kleineres (größeres) Excess Maß für das Prunen des Dichteschätzer Basierten Dendrogramms auswählt. In diesem Sinne kann man die Datenstruktur durch Betrachtung von verschiedenen Versionen des Visualisierungsbaums vollständig untersuchen, insbesondere wenn man die Daten interaktiv im Visualisierungsbaum darstellt.

Methode2

Hier wird die Grundidee für die Veranschaulichung von $\hat{E}_h(N)$ in Abbildung 3.27 (vgl. Abs. 3.3) auf den multivariaten Fall erweitert. Wie in Abschnitt 3.3 erwähnt kann die Modalstruktur der Daten beim Dichteschätzer basierten Clustering dadurch bestimmt werden, indem man den entsprechenden Dichteschätzer Basierten Baum B zerlegt. Wir verwenden hier B_i, $i = 1, ..., k$ für die den k High Density Regions aus dem Dichteschätzer basierten Clustering entsprechenden Subbäume von B, und B_{rr} für $B \backslash (B_1 \cup B_2 \cup, ..., \cup B_k)$.

Im Folgenden wird **Methode2** anhand des Beispiels 3.1.1 vorgestellt, wobei das Single Linkage Verfahren auf Kerndichteschätzer $\hat{f}$ mit LSCV-Bandbreiten basiert. Man prunt das entsprechende Dichteschätzer Basierte Dendrogramm mit den 4 größten intuitiven Maßen (in diesem Fall ist es gleich wie mit den 4 größten Runt Excess Maßen) und erhält jeweils 2, 3, 4 und 5 High Density Cluster. Abbildung 4.29-4.32 zeigt jeweils B_{rr_k} und B_i, $i = 1, ..., k$ für $k = 2, ..., 5$ im Scatterplot (Grafik links) und im **Methode2** (Grafik rechts). Man stellt **Methode2** anhand der Abbildung 4.29 ($k = 2$) wie folgt vor (auch vgl. Abs. 3.3):

- B_1 und B_2 werden in rot und grün eingefärbt;
- B_{rr_2} wird in leicht grau eingefärbt;
- Um die Überlappung bei der grafischen Darstellung zu vermeiden, ersetzt man diejenige Kanten beim Konstruieren des Dichteschätzer basierten Dendrogramms, die einen Punkt mit einer Datengruppe bzw. zwei Datengruppen verbinden, durch die Kanten mit minimaler euklidischen Länge zwischen dem Punkt und der Datengruppe bzw. zwischen den zwei Datengruppen (vgl. Tabelle 3.4), und verändert B entsprechend;
- Die X-Achse im **Methode2** besteht aus den Kanten von dem veränderten B, d.h., der veränderte B wird Kante für Kante in die X-Achse im **Methode2** gelegt;
- Die Y-Achse im **Methode2** steht für die Funktionswerte von $\hat{f}$;
- B_1 hat 112 Daten und 111 Kanten. In diesem Fall werden die 111 Querschnitte auf Basis der 111 Kanten auf die 2 dimensionale Ebene in der mittleren Grafik im **Methode2** gelegt. Die Fläche jedes Querschnitts kann durch die Summe der

Flächen von einer Reihe der Trapeze approximiert werden. Es gilt analog für B_2 und B_{rr_2}.

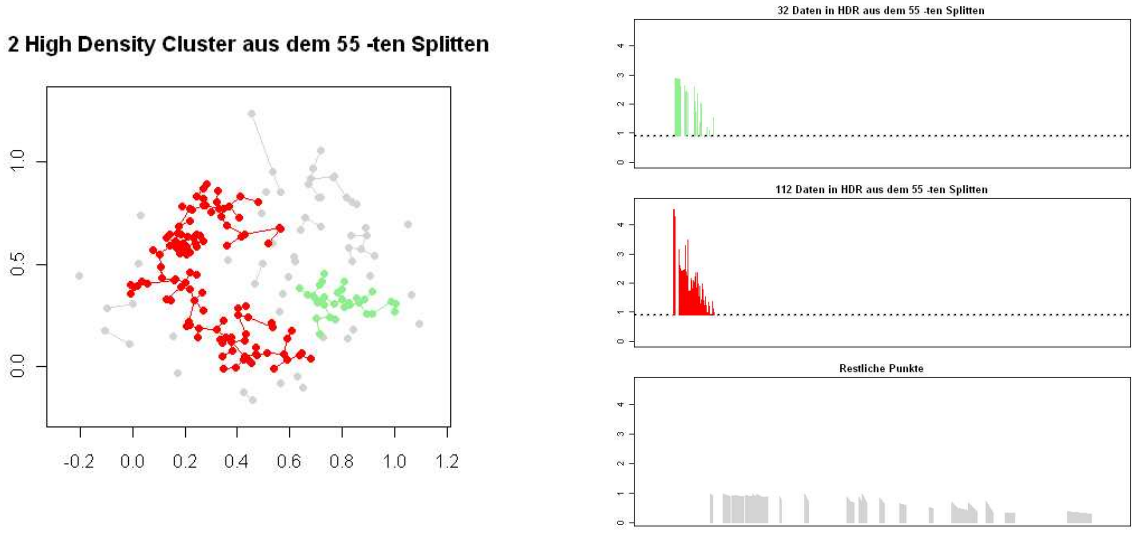

Abbildung 4.29: 2 High Density Regions in Beispiel 3.1.1

Man sieht in Abbildung 4.29-4.32, dass die Form eines High Density Clusters auf diese Art und Weise grafisch dargestellt werden kann. Dadurch hat man die Möglichkeit, die Clusterstruktur der Daten anhand der Grafik zu untersuchen. Beim Prunen des Dichteschätzer basierten Single Linkage Dendrogramms mit dem intuitiven Maß in (3.27) aus dem Ellenbogen-Kriterium erhält man ein 3-Cluster Modell, was mit dem Resultat durch Nutzung des Runt Excess Maßes übereinstimmt (vgl. Abs. 3.2). Abbildung 4.33 zeigt die 10 größten intuitiven Maße aus dem Dichteschätzer basierten Single Linkage Dendrogramm. Die Grafiken in Abbildung 4.29-4.32 deuten klar auf eine 3-Cluster-Struktur aber auch auf eine 4-Cluster-Struktur der Daten in Beispiel 3.1.1 hin.

Im Folgenden zeigt man die Anwendung von **Methode2** in Beispiel 4.1.1. Abbildung 4.34 zeigt die 20 größten Runt Excess Maße (Grafik links) und die 20 größten intuitiven Maße (Grafik rechts) beim Dichteschätzer (Kerndichteschätzer mit $h = (0,5; 0,5; 0,5; 0,5; 0,5)^T$) basierten Single Linkage Verfahren. Mit dem Runt Excess Maß $= 7,6$ oder Intuitive Maß $= 0,02$ aus dem Quasi-Ellenbogen-Kriterium erhält man durch Pruning des Dichteschätzer basierten Single Linkage Dendrogramms 4 High Density Cluster. In Abbildung 4.35-4.37 zeigt man B_{rr_k} und B_i, $i = 1, ..., k$ für $k = 2, ..., 7$ im **Methode2**. Man sieht in Abbildung 4.35-4.37, dass diese Clusterstruktur mit 4 High Density Cluster problematisch ist, weil es noch eine Menge der restlichen Punkte gibt, die aber nicht alle in der Low-Density-Region liegen.

Eigentlich soll die Clusterstruktur der simulierten Daten in Beispiel 4.1.1 einfach zu entdecken sein, wenn man die dahinter liegende gemischte Normalverteilung betrachtet.

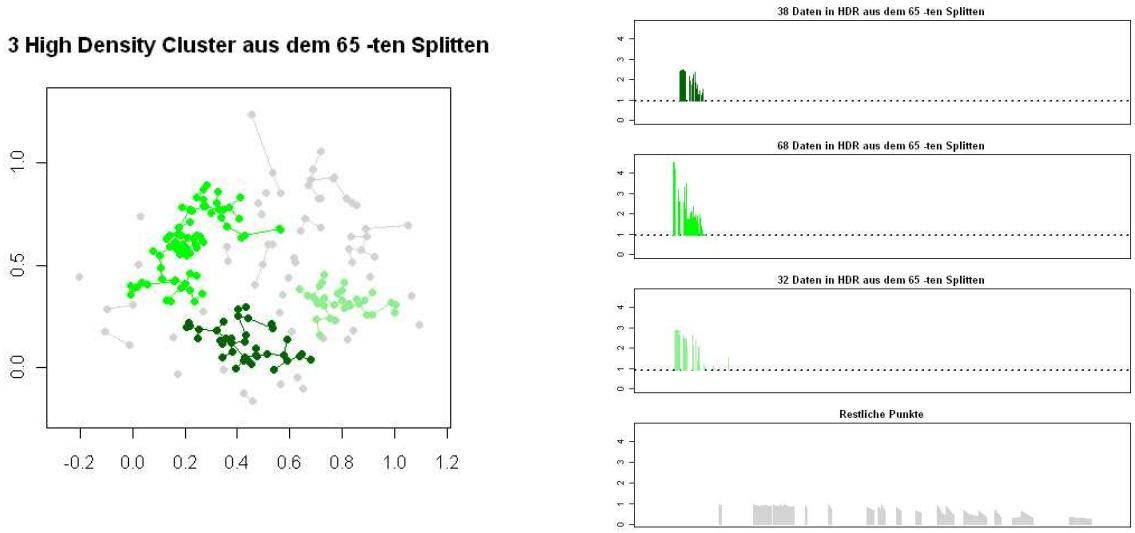

Abbildung 4.30: 3 High Density Regions in Beispiel 3.1.1

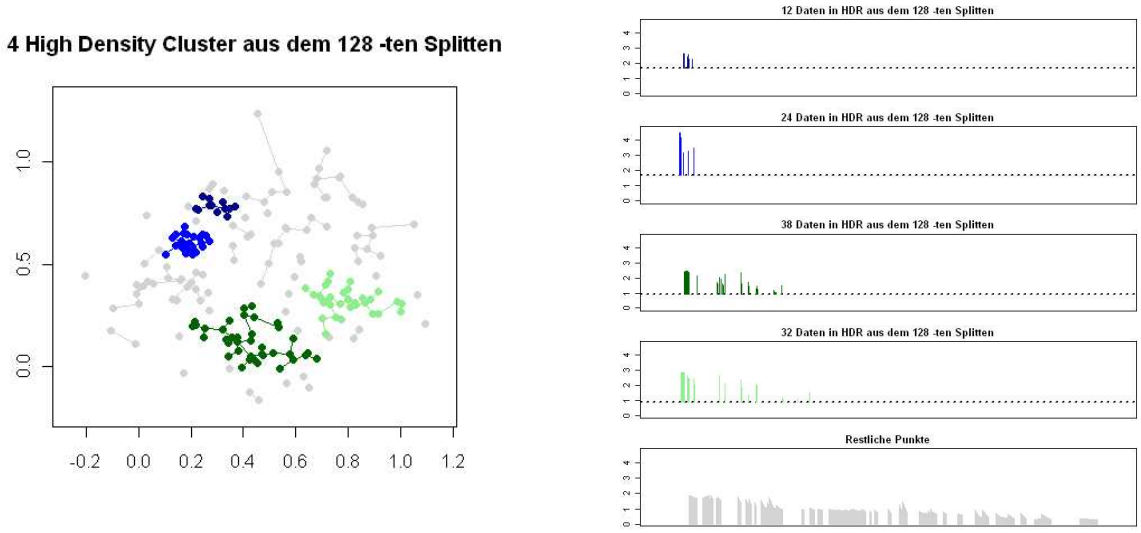

Abbildung 4.31: 4 High Density Regions in Beispiel 3.1.1

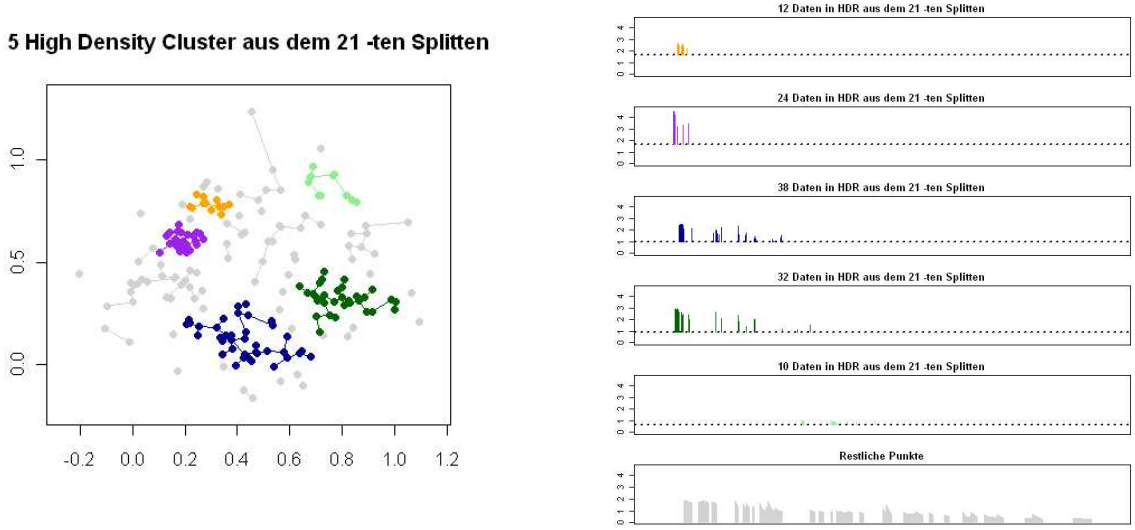

Abbildung 4.32: 5 High Density Regions in Beispiel 3.1.1

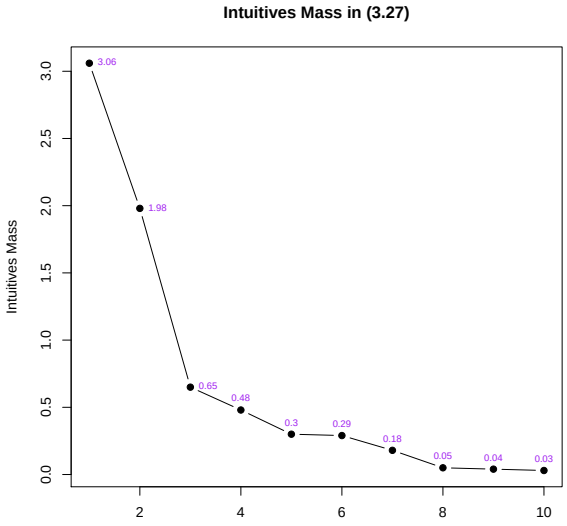

Abbildung 4.33: 10 größte intuitive Maße aus dem Dichteschätzer basierten Single Linkage Dendrogramm in Beispiel 3.1.1

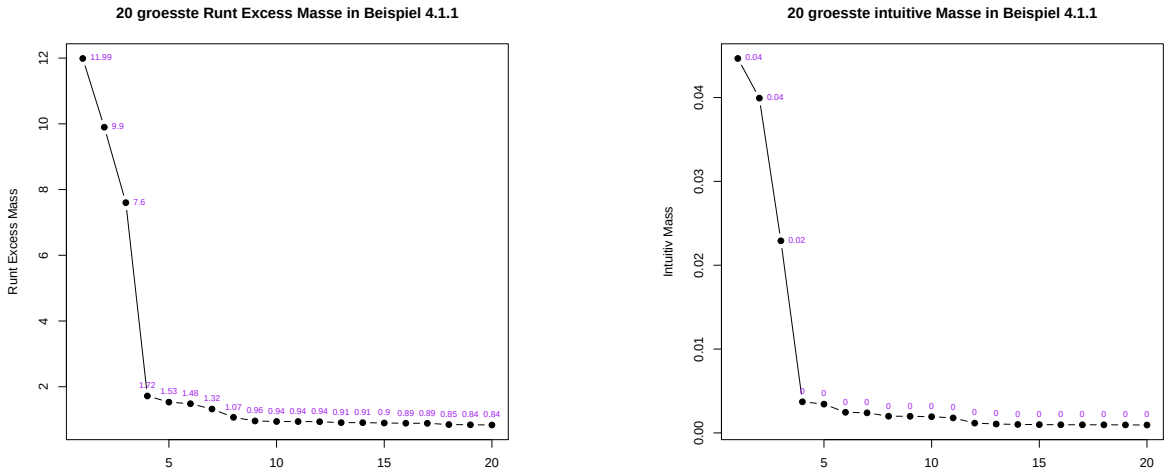

Abbildung 4.34: 20 größte Runt Excess Maße (Grafik links) und 20 größte intuitiven Maße (Grafik rechts) beim Dichteschätzer (Kerndichteschätzer mit $h = (0,5; 0,5; 0,5; 0,5; 0,5)^T$) basierten Single Linkage Verfahren in Beispiel 4.1.1

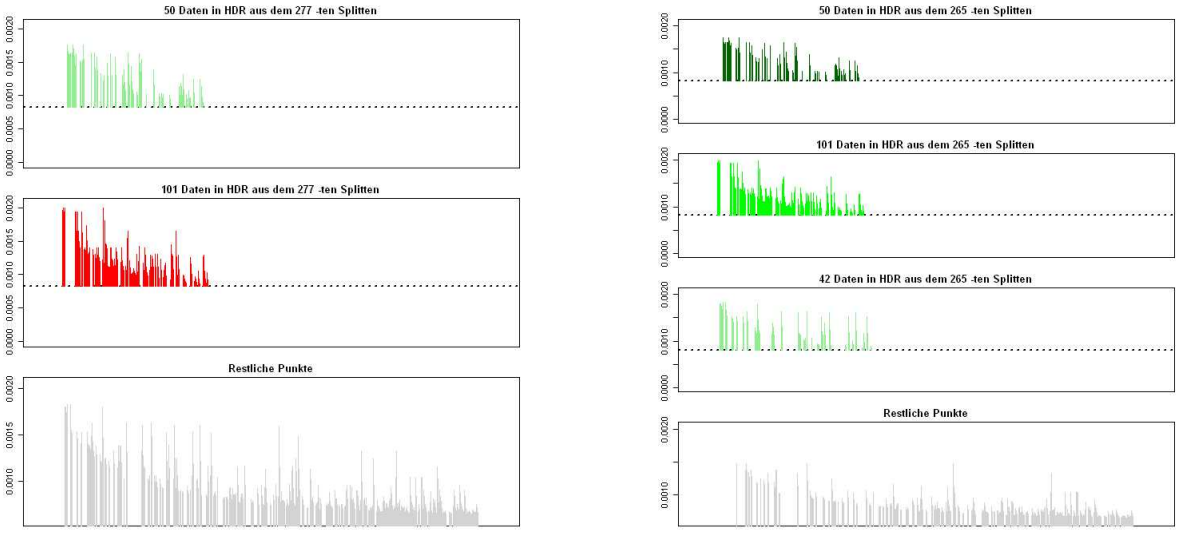

Abbildung 4.35: 2 High Density Cluster im **Methode2** (Grafik links) und 3 High Density Cluster im **Methode2** (Grafik rechts) in Beispiel 4.1.1

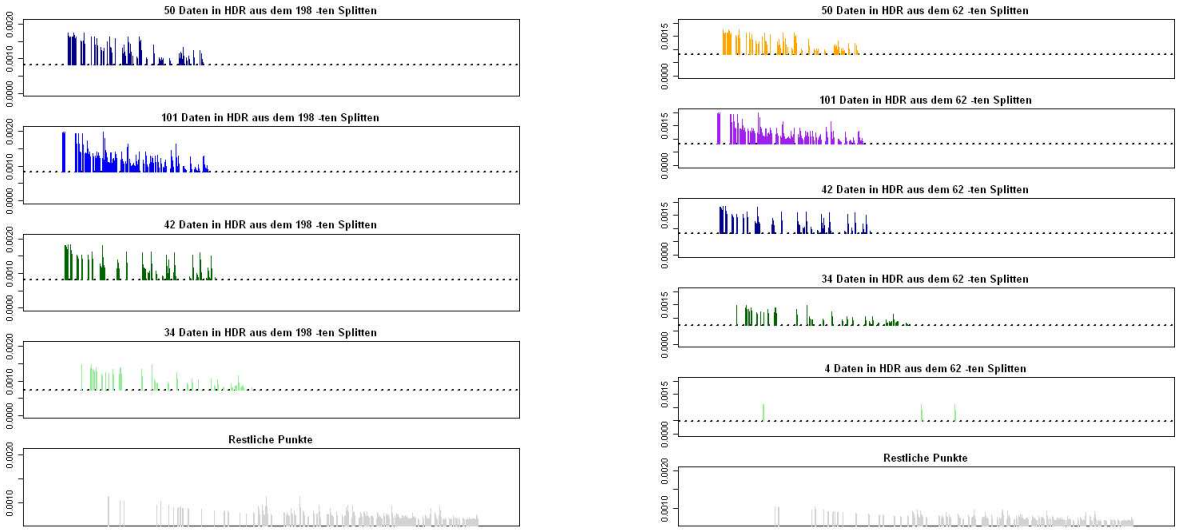

Abbildung 4.36: 4 High Density Cluster im **Methode2** (Grafik links) und 5 High Density Cluster im **Methode2** (Grafik rechts) in Beispiel 4.1.1

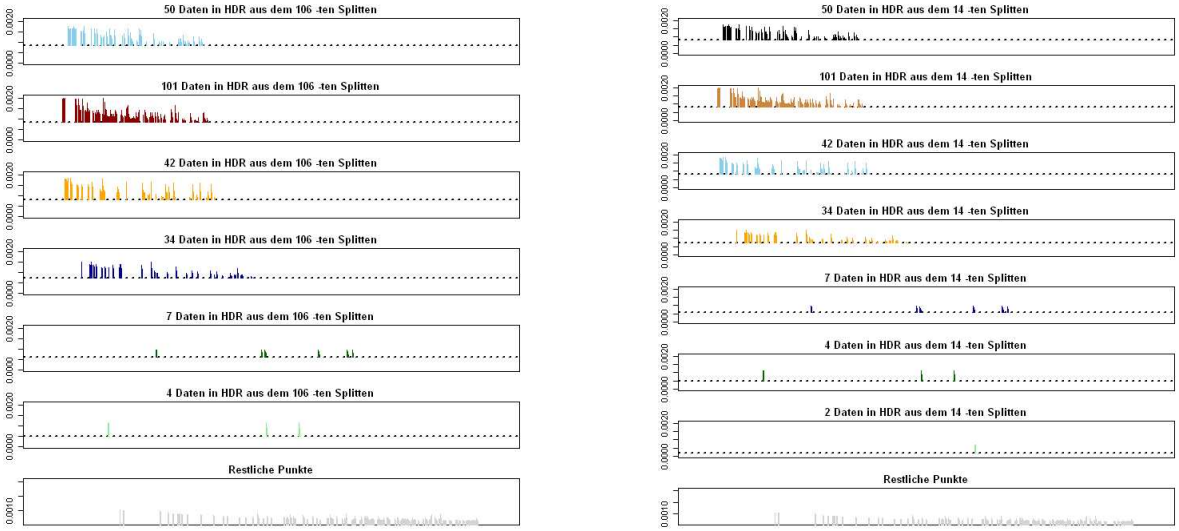

Abbildung 4.37: 6 High Density Cluster im **Methode2** (Grafik links) und 7 High Density Cluster im **Methode2** (Grafik rechts) in Beispiel 4.1.1

Der Grund für das obige Scheitern liegt darin, dass ein gewisses Teil des Wahrscheinlichkeitsmaßes der gesuchten Cluster unter dem „max/max/max“ Dichteniveau liegt (vgl. Abs. 3.3), wenn man das Single Linkage Verfahren auf Basis des Kerndichteschätzers mit $h = (0,5; 0,5; 0,5; 0,5; 0,5)^T$ verwendet. In diesem Fall ist die wahre Clusterstruktur gut zu identifizieren, wenn man das Complete Linkage Verfahren auf Basis des Kerndichteschätzers mit $h = (0,5; 0,5; 0,5; 0,5; 0,5)^T$ benutzt. Abbildung 4.38 zeigt die 20 größten Runt Excess Maße (Grafik links) und die 20 größten intuitiven Maße (Grafik rechts) beim Dichteschätzer basierten Complete Linkage Verfahren. Mit dem Runt Excess Maß

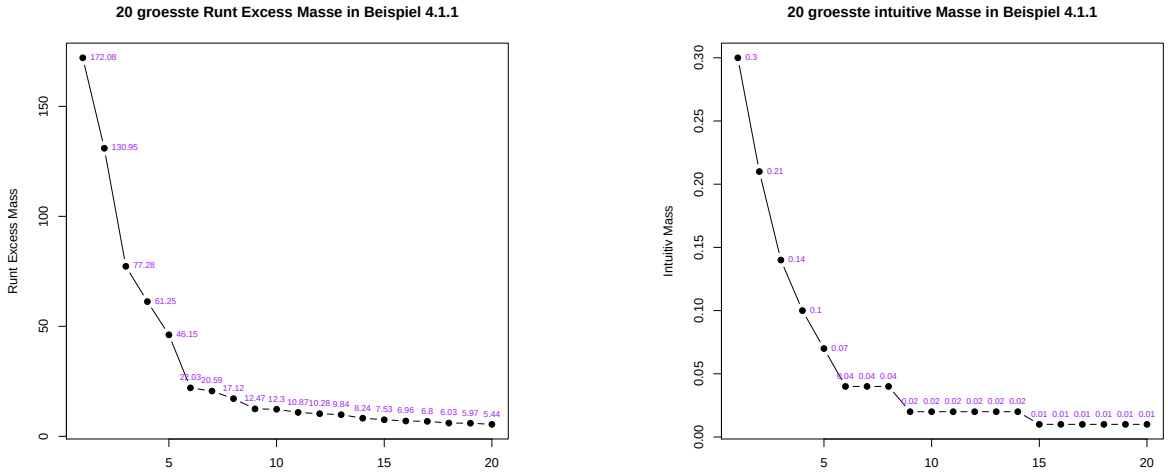

Abbildung 4.38: 20 größte Runt Excess Maße (Grafik links) und 20 größte intuitiven Maße (Grafik rechts) beim Dichteschätzer (Kerndichteschätzer mit $h = (0,5; 0,5; 0,5; 0,5; 0,5)^T$) basierten Complete Linkage Verfahren in Beispiel 4.1.1

$= 46,15$ oder Intuitive Maß $= 0,07$ aus dem Quasi-Ellenbogen-Kriterium erhält man durch Pruning des Dichteschätzer basierten Complete Linkage Dendrogramms 6 High Density Cluster. Man zeigt die 6 High Density Cluster im **Methode2** in Abbildung 4.39. In Abbildung 4.39 sieht man, dass 487 Punkte in den 6 High Density Clustern aufgeteilt worden sind, was der wahren Datenstruktur entspricht.

In der Tat spielt die Visualisierung der High Density Cluster beim Dichteschätzer basierten hierarchischen Verfahren eine wichtige Rolle, weil es in der explorativen Datenanalyse kaum ein optimales Pruning des entsprechenden Dendrogramms geben kann, es sei denn, dass die Daten eine ganz klare Clusterstruktur besitzen. In diesem Sinne gibt es eine gewisse Gemeinsamkeit zwischen der Festlegung eines geeigneten Maßes für das Pruning des Dichteschätzer basierten Dendrogramms und der Bestimmung eines optimalen Glättungsparameters bei Kerndichteschätzung. Die Zusammenarbeit von dem intuitiven Maß und **Methode2** hilft beim Anwenden des Dichteschätzer basierten Clusterings in

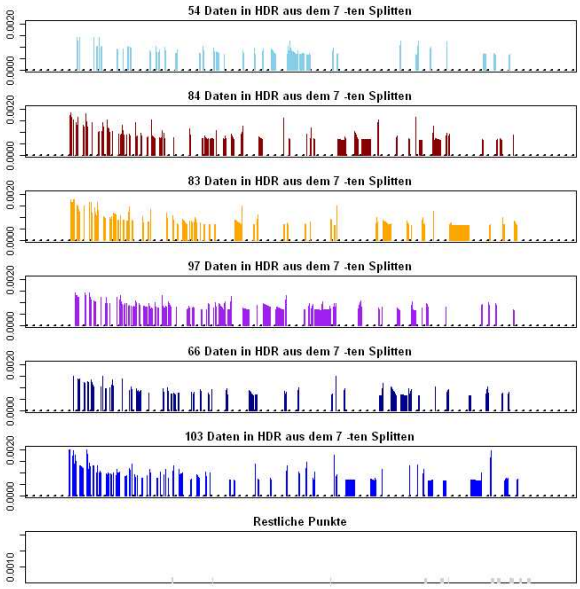

Abbildung 4.39: 6 High Density Cluster aus Prunen des Dichteschätzer basierten Complete Linkage Dendrogramms im **Methode2** in Beispiel 4.1.1

der explorativen Datenanalyse, um alle möglichen signifikanten High Density Cluster zu identifizieren und keine Information über die Clusterstruktur der Daten zu verlieren.

Zum Zweck der Datenexploration ist stark zu empfehlen, die Daten mit verschiedenen grafischen Methoden zu visualisieren, weil jede Methode eigene Vor- und Nachteile hat und in der Regel nur gewisse Facetten der Daten darstellen kann. Z.B., ein Visualisierungsbaum zeigt die Clusterstruktur der Daten aber nicht die Form der High Density Cluster und die entsprechenden Dichteniveaus, die aber im **Methode2** gut darstellbar sind. In Zukunft ist neue interaktive Visualisierungs-Software für die Unterstützung der Anwendung der Dichteschätzung in der explorativen Datenanalyse zu entwickeln aufgrund der folgenden Tatsachen:

- Es besteht eine starke Abhängigkeit von Glättungsparametern, Fest-Kerndichteschätzern und darauf basierten statistischen Modellen;
- Durch verschiedene Prunings eines Dichteschätzer basierten Dendrogramms erhält man unterschiedliche Versionen der Daten;
- Es wurde in diesem Kapitel gezeigt, dass eine interaktive Darstellung der Information aus Kerndichteschätzung bzw. den darauf basierten statistischen Verfahren beim Untersuchen der Datenstruktur hilft;
- Interaktive Softwares für die explorative Datenanalyse wurden bereits gut entwickelt (vgl. Softwares am ROSUDA der Universität Augsburg unter http://rosuda.org/software/).

5 Kerndichteschätzung in R

R ist eine objekt-orientierte und interpretierte Sprache und Programmierumgebung für Statistical Computing, die ursprünglich (1994) von Ross Ihaka und Robert Gentleman an der University of Auckland, Neuseeland entwickelt worden ist. **R** ist eine Open Source Software basiert auf den Definitionen von **S**, die von John Chambers und seinen Kollegen in den Bell Laboratories, New Jersey seit den 70er Jahren für Probleme der Datenanalyse entwickelt wurde. In 1998 vergab die ACM (Association for Computing Machinery) ihren „Software System Award“ für **S** (vgl. Vorwort in Ligges (2004)):

> . . . the S system, which has forever altered the way people analyze, visualize, and manipulate data . . . S is an elegant, widely accepted, and enduring software system, with conceptual integrity, thanks to the insight, taste, and effort of John Chambers.

Für eine ausführliche Vorstellung von **R** verweisen wir auf die Arbeit von Venables & Ripley (2002), Ligges (2004), Chambers (2008). Nähere Information über **R** befindet sich auch unter http://r-project.org/.

R kann über Pakete erweitert werden, die zusätzliche Funktionalität bereitstellen. In den letzten Jahren wurde eine Vielzahl von **R** Paketen in verschiedensten Bereichen der Datenanalyse entwickelt. Da **R**-Funktionen in diese Arbeit für die Unterstützung der Argumente intensiv eingesetzt wurden, werden in diesem Kapitel die **R**-Funktionen für Kerndichteschätzung anhand von Beispielen erläutert. Das vorliegende Kapitel gliedert sich in 2 Teile. Im ersten Teil werden die mit Kerndichteschätzung relevanten **R**-Pakete bzw. Funktionen anhand der **Hidalgo** und **geyser** Daten vorgestellt. Im zweiten Teil werden zwei aus der Nutzung der **R** Funktionen in dieser Arbeit entstehenden Probleme besprochen und ein paar Bemerkungen zu den im ersten Teil vorgestellten **R** Funktionen gegeben.

5.1 Kerndichteschätzung in R

Für die univariate Kerndichteschätzung steht eine Vielzahl von **R** Funktionen zur Verfügung. Im Unterschied dazu bietet **R** wenige Pakete für Kerndichteschätzung im multivariaten Fall. In diesem Abschnitt werden die mit Kerndichteschätzung relevanten **R** Pakete bzw. Funktionen anhand der **Hidalgo** und **geyser** Daten vorgestellt.

In Tabelle 5.1 werden die **R**-Funktionen für die Bestimmung des Glättungsparameters bei univariater Kerndichteschätzung mit Gaussian Kernfunktion aufgelistet. Die entsprechenden Bandbreiten für die **Hidalgo** Daten werden in der letzten Spalte der Tabelle

gezeigt. In Tabelle 5.1 stellt man diejenige Terme in Schrägschrift dar, die noch zu klären sind.

Nr.	Funktionsname	Paket	Methode	Bandbreite
1	*bandwidth.nrd*[1]	MASS	Rule of Thumb	0,0184
2	*bcv*[2]	MASS	BCV	0,0146
3	bw.nrd0	stats	Rule of Thumb	0,0039
4	bw.nrd	stats	Normal-Reference	0,0046
5	*bw.ucv*[3]	stats	LSCV	0,0005
6	bw.bcv	stats	BCV	0,0037
7	bw.SJ	stats	Plug-In (Sheather & Jones (1991))	0,0012
8	dpih	KernSmooth	Direct Plug-In für Histogramm (Wand (1995))	0,0029
9	dpik	KernSmooth	Direct Plug-In (Sheather & Jones (1991))	0,0020
10	*hcv*[4]	sm	LSCV	
11	hnorm	sm	Normal-Reference	0,0046
12	hsj	sm	Direct Plug-In (Sheather & Jones (1991))	0,0000
13	*npudensbw*[5]	np	LCV (Default), LSCV, Normal-Reference	0,0000
14	plugin.density	plugdensity	Plug-In (Engel et al. (1994))	0,0012
15	*ucv*[6]	MASS	LSCV	0,0021
16	*width.SJ*[7]	MASS	Plug-In (Sheather & Jones (1991))	0,0048

Tabelle 5.1: Univariate Bandbreitenselektoren in **R**

Erklärungen zu Tabelle 5.1:

1. bandwidth.nrd(MASS) liefert einen Glättungsparameter von $4h_{nr}$, wobei h_{nr} für die Normal-Reference Bandbreite steht;

2. bcv(MASS) liefert einen Glättungsparameter von $4h_{bcv}$, wobei h_{bcv} für die BCV-optimale Bandbreite steht;

3. Wegen der Bindungen in den **Hidalgo** Daten gibt es eine triviale optimale LSCV-Bandbreite, nämlich $h_{lscv} = 0$ (vgl. Abs. 2.3). bw.ucv(stats) liefert in diesem Fall einen Glättungsparameter am linken Rand des gesuchten Bereichs;

4. Man bekommt eine Fehlermeldung wegen der Bindungen in den **Hidalgo** Daten;

5. Bei npudensbw(np) stehen drei Bandbreitenselektoren (LCV (Default), LSCV und Normal-Reference) zur Verfügung. Der LCV-Bandbreitenselektor wird hier benutzt und liefert eine ziemlich kleine Bandbreite wegen der Bindungen in den **Hidalgo** Daten;

6. ucv(MASS) liefert einen Glättungsparameter von $4h_{lscv}$, wobei h_{lscv} für die LSCV-optimale Bandbreite steht;

7. width.SJ(Mass) liefert einen Glättungsparameter von $4h_{sj}$, wobei h_{sj} für die Plug-In-optimale Bandbreite steht.

Tabelle 5.2 zeigt die multivariaten Bandbreitenselektoren in **R**. Man sieht in Tabelle 5.2, dass die meisten Funktionen aus Paket **ks** kommen. Das **R** Paket **ks** bietet auch

Funktionen, mit denen man nach Duong (2004) eine optimale volle Bandbreitenmatrix berechnen kann. Im Unterschied dazu liefern die anderen **R** Funktionen eine diagonale Bandbreitenmatrix für die multivariate Kerndichteschätzung. Ein paar Erklärungen zu

Nr.	Funktionsname	Paket	Dimension der Daten	Methode
10	*hcv*[1]	sm	1-2	LSCV
17	Hbcv	ks	2-6	BCV für volle Bandbreitenmatrix (Duong (2004))
18	Hbcv.diag	ks	2-6	BCV
11	hnorm	sm	1-3	Normal-Reference
19	Hlscv	ks	2-6	LSCV für volle Bandbreitenmatrix (Duong (2004))
20	Hlscv.diag	ks	2-6	LSCV
21	Hpi	ks	2-6	Plug-In (Duong (2004))
22	Hpi.diag	ks	2-6	Plug-In (Duong (2004))
23	Hscv	ks	2-6	SCV
24	Hscv.diag	ks	2-6	SCV für volle Bandbreitenmatrix (Duong (2004))
13	*npudensbw*[2]	np	1 -	LCV, LSCV, Normal-Reference

Tabelle 5.2: Multivariate Bandbreitenselektoren in **R**

Tabelle 5.2:

1. hcv(sm) liefert nicht immer einen LSCV-optimalen Glättungsparameter. Ein Beispiel befindet sich in Abschnitt 5.2;

2. Im multivariaten Fall (Dimension der Daten größer 6) funktioniert npudensbw(np) mit LCV und LSCV Bandbreitenselektoren in der Regel nicht gut.

Die Zuordnung der Bandbreitenselektoren in **R** nach **R** Paketen bzw. nach der dahinter stehenden Methodik wird in Abbildung 5.1 veranschaulicht, wobei die Abkürzung (unten im Rechteck) NR bzw. RT für „Normal-Reference“ bzw. „Rule of Thumb“ steht. Die Nummer im Kreis steht für die entsprechende **R** Funktion (vgl. Tabelle 5.1 und 5.2).

Tabelle 5.3 zeigt die **R** Funktionen für die Berechnung des Kerndichteschätzers.

Funktionsname	Paket	Bandbreite (Default)	Dim.	Methode
akj	quantreg	Normal-Reference	1	Adaptive Kerndichteschätzung
bkde	KernSmooth	Oversmoothed Bandbreite	1	Binned Kerndichteschätzer
bkde2D	KernSmooth	Muss vorgegeben werden	2	Binned Kerndichteschätzer
density	stats	Normal-Reference	1	Binned Kerndichteschätzer
kde[1]	ks	Muss vorgegeben werden	1-6	Binned und Nicht-Binned Kerndichteschätzer
plugin.density	plugdensity	Bandbreite (Engel et al. (1994))	1	Nicht-Binned Kerndichteschätzer
npudens[2,3]	np	aus npudensbw(np)	1-	Methode von Li & Racine (2003)
sm.density	sm	Normal-Reference	1-3	Nicht-Binned Kerndichteschätzer

Tabelle 5.3: Kerndichteschätzer in **R**

Erklärungen zu Tabelle 5.3:

1. Mit kde(ks) kann man Kerndichteschätzer für 1-6 dimensionale Daten berechnen. Die Funktion liefert auch Binned Kerndichteschätzer für 1-4 dimensionale Daten. Beim Nutzen der Funktion kde(ks) muss der Glättungsparameter vorgegeben werden;

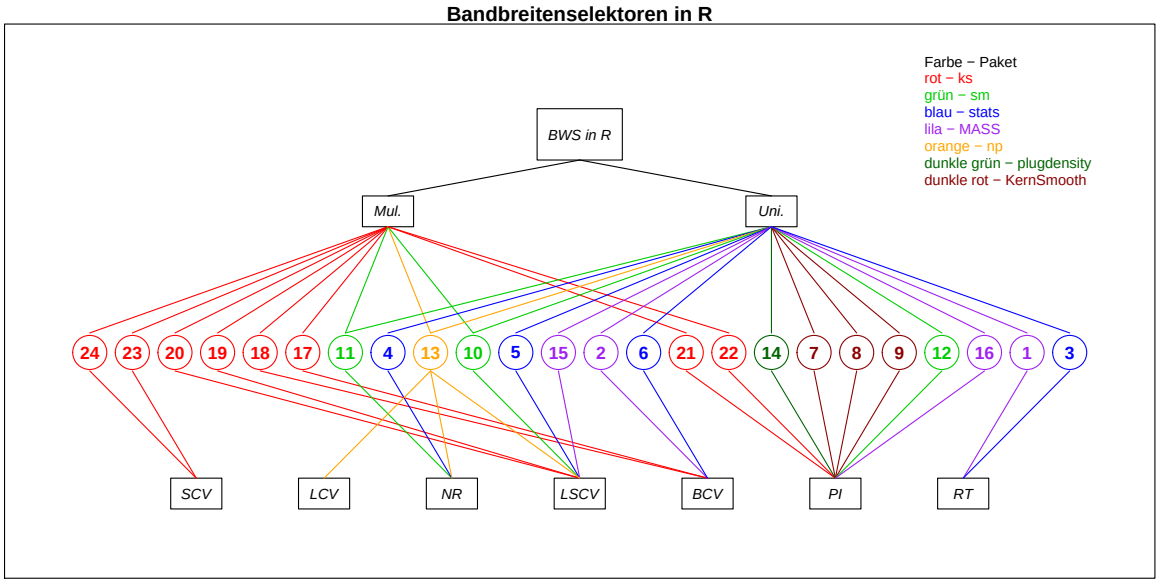

Abbildung 5.1: Veranschaulichung der Zuordnung der Bandbreitenselektoren in **R**

2. Mit npudens(np) kann man Kerndichteschätzer für Daten berechnen, die diskrete Variable enthalten. Die entsprechenden statistischen Modelle befinden sich in Li & Racine (2003, 2007);

3. Mit npudens(np) kann man Kerndichteschätzer für Daten mit beliebiger Dimension berechnen. Zu bemerken ist, dass man beim Nutzen dieser Funktion für multivariate Kerndichteschätzung den Glättungsparameter angeben sollte, weil ansonsten npudens(np) die Bandbreite aus npudensbw(np) mit der LCV (Default) Methode nehmen würde und dies im multivariaten Fall (Dimension der Daten größer 6) in der Regel nicht gut funktioniert.

Zum Schluss dieses Abschnitts ist das **R** Paket **feature** zu erwähnen, das auf der Arbeit von Duong et al. (2007) beruht, deren Grundidee bereits in Abschnitt 3.2 dieser Arbeit kurz vorgestellt wurde. Hier zeigt man die **R** Funktion featureSignif(feature) anhand der **geyser** Daten. Abbildung 5.2 zeigt die **geyser** Daten im Feature Plot. Die für die Darstellung eines Feature Plots wichtigen Parameter in featureSignif(feature) sind (vgl. Abbildung 5.2):

- *bw*: Bandbreite für den Kerndichteschätzer. Wenn *bw* nicht vorgegeben ist, dann wird eine Reihe von möglichen Glättungsparametern für die Kerndichteschätzung verwendet, in welchem Fall man eine interaktive Grafik bekommt, wobei die Auswahl des Glättungsparameters durch den Slider unten in der Grafik kontrolliert wird. In unserem Beispiel ist *bw* nicht vorgegeben;

- *addData*: Wenn TRUE, dann werden die Datenpunkte zum Feature Plot hinzugefügt. In unserem Beispiel wird der default Wert von *addData* (FALSE) genommen;
- *addSignifGradRegion*: Wenn TRUE, dann werden die Regionen mit signifikanter $\hat{f}'$ (vgl. Abs. 3.2) gezeigt. Der default Wert ist FALSE. In unserem Beispiel wird TRUE genommen;
- *addSignifCurvRegion*: Wenn TRUE, dann werden die Regionen mit signifikanter $\hat{f}''$ (vgl. Abs. 3.2) gezeigt. Der default Wert ist FALSE. In unserem Beispiel wird TRUE genommen;
- *gradCol*: Farbe für die Darstellung der Regionen mit signifikanter $\hat{f}'$. Der default Wert ist *grenn*4, welcher in unserem Beispiel verwendet wird;
- *curvCol*: Farbe für die Darstellung der Regionen mit signifikanter $\hat{f}''$. Der default Wert ist *blue*, welcher in unserem Beispiel verwendet wird.

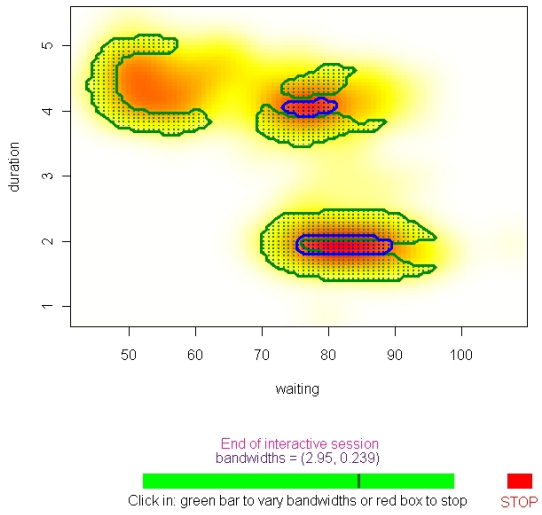

Abbildung 5.2: **geyser** Daten im Feature Plot

5.2 Probleme der R Funktionen bei Kerndichteschätzung

Bei Nutzung der obigen **R** Funktionen für die Unterstützung der Argumente in dieser Arbeit sind ein paar Probleme entstanden. Im Folgenden werden zwei dieser Probleme anhand von Beispielen diskutiert.

Randproblem bei Kerndichteschätzung

Ein triviales Problem der Kerndichteschätzung in **R** liegt in der Behandlung des Randproblems. Das Randproblem bei Kerndichteschätzung kann wie folgt beschrieben werden:

- Der Definitionsbereich von manchen Variablen ist beschränkt;
- Die wahre Dichte ist nicht kontinuierlich am Rand, nämlich, der Funktionswert ist positiv im Definitionsbereich und gleich Null außerhalb des Definitionsbereichs;
- Mit einer gewöhnlichen symmetrischen Kernfunktion wird der Kerndichteschätzer stetig über den Rand hinaus gesetzt, wenn man die unbekannte Dichte an einem Punkt im Randbereich schätzt. Dies führt zum Bias im Randbereich.

Das Randproblem wird bei multivariater Kerndichteschätzung noch auffälliger, weil der Randbereich mit der Erhöhung der Datendimension größer wird. In den letzten Jahren wurde eine Vielzahl von statistischen Verfahren vorgeschlagen, um dem Randproblem bei Kerndichteschätzung zu begegnen. Typische Methoden sind

- Spiegelungsverfahren (Schuster (1985));
- Nutzung einer Randkernfunktion im Randbereich (Müller (1991), Jones (1993), Dong & Simonoff (1994), Jones & Foster (1996), Müller & Stadtmüller (1999), usw.);
- Nutzung einer nichtsymmetrischen Kernfunktion (Chen (2000), Scaillet (2003), Bouezmarni & Rombouts (2007).

Als Beispiel zeigt Abbildung 5.3 zwei Kerndichteschätzer mit der Biweight Kernfunktion und $h = 0,5$ für die Daten in Beispiel 1.2.1, wobei die linke (rechte) Grafik den Kerndichteschätzer ohne (mit) Behandlung des Randproblems zeigt (vgl. Abs. 1.2). Man sieht in Abbildung 5.3, dass der Kerndichteschätzer in der linken Grafik am linken Rand problematisch ist, weil das Tempo der Bewegung auf keinen Fall kleiner Null ist und deswegen die wahre Dichtefunktion am linken Rand eigentlich nicht kontinuierlich sein soll.

Problem der Bandbreitenselektoren

Das Hauptproblem bei Bandbreitenselektoren in **R** liegt darin, dass die **R** Bandbreitenselektoren nicht immer das gesuchte lokale Minimum der Zielfunktion liefern, was in der praktischen Datenanalyse zum Irrtum führen kann. Anhand der folgenden zwei kleinen Beispiele wird dieses Problem veranschaulicht.

Beispiel 5.1.1

Man fügt Störungsterme $\rho_i \sim U(-0,00005; 0,00005)$, $i = 1, ..., 485$ zu den **Hidalgo** Daten hinzu und bestimmt $\hat{h}_{lscv}$ mit **R** Funktionen bw.ucv(stats), hcv(sm) und npudensbw(np) wie folgt:

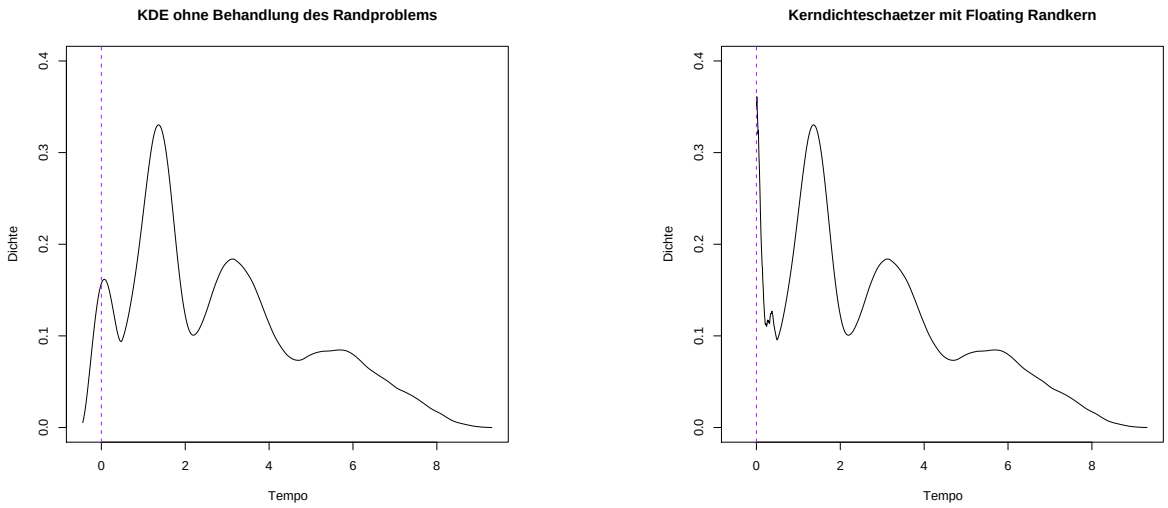

Abbildung 5.3: Kerndichteschätzer für die Daten in Beispiel 1.2.1 ohne (Grafik links) und mit (Grafik rechts) Randproblem-Behandlung

```
st.bwucv<-bw.ucv(thickn,lower=1e-10,nb=1000)
st.hcv<-hcv(thickn,hstart=1e-20,ngrid=101)
st.npu<-npudensbw(thickn,bwmethod="cv.ls")$bw
```

Als Rückgabe bekommt man $st.hcv = 2,2904e{-}05$, $stnpu = 2,3160e{-}05$ und $st.bwucv = 1,0496e-10$ mit der folgenden Warnung:

```
Warning message:
In bw.ucv(thickn[[i]], lower = 1e-10) :
Minimum trat am Rand der Bereichs auf
```

In der Tat liegt $\hat{h}_{lscv}$ ca. an $2.3e-05$, wie der Funktionsverlauf von $LSCV(h)$ im Bereich von $h = [1,0e-05; 4,5e-05]$ in Abbildung 5.4 zeigt. In diesem Sinne liefert bw.ucv(stats) kein sinnvolles Resultat. Der Grund liegt darin, dass die binned Methode dabei verwendet wird und damit die Modifikation der Daten (mit ρ_i) den Funktionsverlauf von $LSCV$ nicht verändert, wenn man den default Wert von nb ($nb = 1000$) nimmt. Falls man aber z.B. $nb = 1e+5$ verwendet, dann liefert bw.ucv(stats) eine Bandbreite $\hat{h}_{lscv} = 2,3185e-05$, was dem gesuchten lokalen Minimum von $LSCV(h)$ entspricht. Damit dass bw.ucv(stats) in solche Situation ein sinnvolles Resultat liefert, soll nb so gewählt werden, dass die Größenordnung von $range(x)/nb$ gleich oder kleiner der Größenordnung von ρ_i ist, wobei $range(x)$ für die Reichweite der Daten steht.

Beispiel 5.1.2

In diesem Beispiel berechnet man $\hat{h}_{lscv}$ für die **geyser** Daten mit **R** Funktionen Hlscv(ks), hcv(sm) und npudensbw(np). Ohne Modifizierung der Daten hat $LSCV(h)$ ein

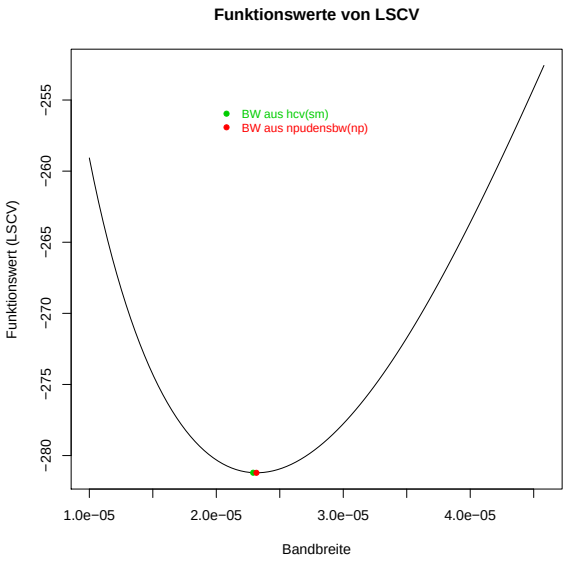

Abbildung 5.4: Funktionsverlauf von $LSCV(h)$ im Bereich von $h = [1,0e-05; 4,5e-05]$ in Beispiel 5.1.1

globales Minimum an $h = (0,0)^T$ (vgl. Abs. 2.3). Die Resultate in diesem Beispiel sind $(0,8297; 0,0686)^T$ aus hcv(sm), $(8,9435e-17; 9,5681e-17)^T$ aus npudensbw(np) und die folgende Fehlermeldung aus Hlscv(ks):

```
Fehler in solve.default(cov, ...) :
System ist für den Rechner singulär:
reziproke Konditionszahl = 4.32152e-18
```

Die Fehlermeldung entsteht, weil die Bandbreitenmatrix H in der 124-ten Iteration nicht invertierbar ist. Es stellt sich nun die Frage, welche der zwei anderen **R** Funktionen die gesuchte $\hat{h}_{lscv}$ liefert. Zuerst stellt man in Abbildung 5.5 die Funktionswerte von $LSCV(h)$ im Bereich von $[0,8297 * 0,9; 0,8297 * 1,1] \times [0,0686 * 0,9; 0,0686 * 1,1]$ im Imageplot (Grafik links) und Contour Plot (Grafik rechts) dar, wobei die Position von $\hat{h}_{lscv}$ aus hcv(sm) mit grünem Kreuz markiert wird. Man sieht in Abbildung 5.5, dass das Resultat aus hcv(sm) $(0,8297; 0,0686)^T$ kein lokales Minimum von $LSCV(h)$ ist. Analog wird das Resultat aus npudensbw(np) in Abbildung 5.6 überprüft. Man sieht in Abbildung 5.6, dass das Resultat aus npudensbw(np) $(8,9435e-17; 9,5681e-17)^T$ eben kein lokales Minimum von $LSCV(h)$ ist.

In Beispiel 5.1.1 und 5.1.2 sieht man, dass man beim Bestimmen von $\hat{h}_{lscv}$ vorsichtig mit **R** Funktionen bw.ucv(stats), hcv(sm) und npudensbw(np) umgehen soll. Es ist in den obigen Situationen nötig, deren Resultate grafisch zu überprüfen und evtl. mit Resultaten aus anderen Funktionen zu vergleichen, damit man einen sinnvollen

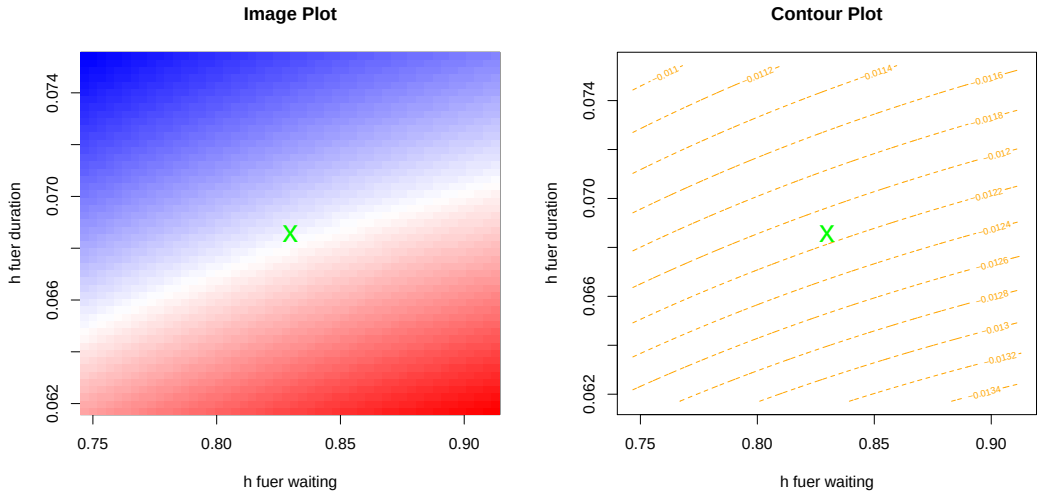

Abbildung 5.5: Funktionsverlauf von $LSCV(h)$ im Bereich von $[0,8297 * 0,9; 0,8297 * 1,1] \times [0,0686 * 0,9; 0,0686 * 1,1]$ in Beispiel 5.2.1

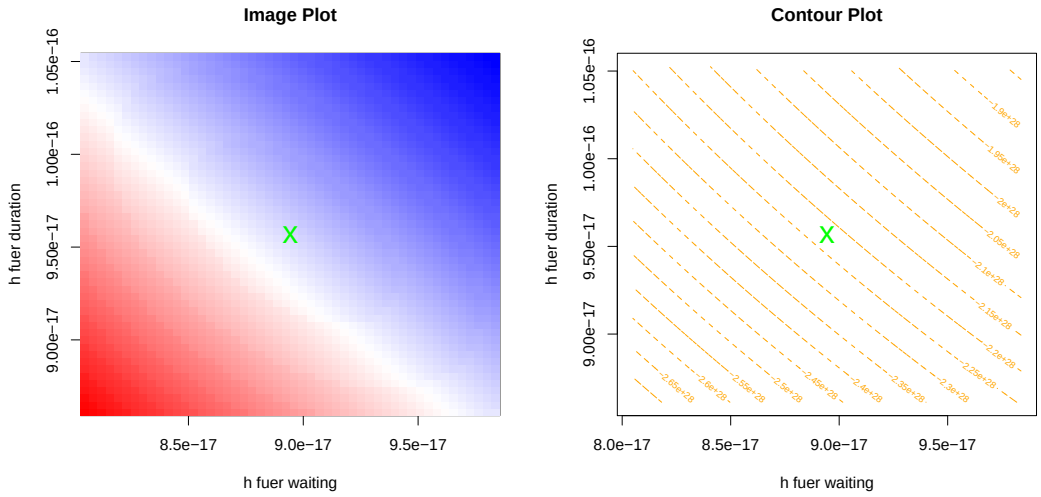

Abbildung 5.6: Funktionsverlauf von $LSCV(h)$ im Bereich von $[8,9435e-17 * 0,9; 8,9435e-17 * 1,1] \times [9,5681e-17 * 0,9; 9,5681e-17 * 1,1]$ in Beispiel 5.2.1

Glättungsparameter erhalten kann. Es ist hilfreich, die Bindungen der Daten vor der Kerndichteschätzung zu untersuchen (vgl. Abs. 2.3). Beim Bestimmen von $\hat{h}_{bcv}$ gibt es auch Problem bei Auswahl eines geeigneten lokalen Minimums, falls $BCV(h)$ mehrere lokale Minima im gesuchten Bereich besitzt. Eine Diskussion über dieses Problem befindet sich in Abschnitt 2.3.

Zum Schluss dieses Abschnitts noch ein paar Bemerkungen zur Kerndichteschätzung in **R**:

- Die **R** Funktionen für die Kerndichteschätzung basieren auf **S3** Methoden, wobei die Struktur eines Objekts nur implizit angegeben werden kann, während sie bei **S4** Methoden genau spezifiziert werden muss. Eine genaue Beschreibung von **S4** Methoden befindet sich in der Arbeit von Chambers (1998), Venables & Ripley (2003) und Chambers (2008);

- Die zurückgelieferten Objekte aus **R** Bandbreitenselektoren sind meist vom Typ *numeric*, während die Objekte aus plugin.density(plugdensity) und npudensbw(np) vom Typ *list* sind, was impliziert, dass der Typ des Resultats aus einem Bandbreitenselektor zu beachten ist, wenn dieses Resultat als Argument für eine andere **R** Funktion verwendet werden soll;

- Alle **R** Funktionen für das Berechnen des Kerndichteschätzers liefern Objekte vom gleichen Typ *list* zurück aber mit verschiedenen Längen und Klassen, was impliziert, dass diese Resultate nicht kompatibel verwendbar sind;

- Es ist manchmal nicht einfach, die Ursache für einen Fehler beim Durchführen einer **R** Funktion herauszufinden, z.B., den Fehler bei Hlscv(ks) in Beispiel 5.1.2. Die zwei Hauptgründe dafür sind:
  - eine **R** Funktion wird in vielen Fällen in mehreren Computersprachen (R, C, Fortran) geschrieben;
  - die Zwischenergebnisse werden in der Regel nicht gezeigt.

6 Zusammenfassung und Ausblick

Die Dichteschätzung gehört zu einem der wichtigsten Werkzeuge in der explorativen Datenanalyse, dessen Ziel darin besteht, die unbekannte Dichte der Daten anhand der vorhandenen empirischen Beobachtungen mittels statistischer Verfahren abzuschätzen und die in den Daten versteckte nützliche Information im Betracht der geschätzten Dichte zu extrahieren. Heutzutage spielt nichtparametrische Dichteschätzung in der explorativen Datenanalyse mit der Entwicklung der Computertechnik eine immer wichtigere Rolle, weil parametrische Dichteschätzung stark von der Modellannahme abhängt und in der Praxis oft wegen des Mangels der a priori Information nicht einwandfrei einsetzbar ist. Die vorliegende Arbeit konzentriert sich auf eine in der Praxis oft eingesetzte Familie der nichtparametrischen Dichteschätzung, nämlich die Kerndichteschätzung. Ziel dieser Arbeit ist es, zu untersuchen, wie die unbekannte Struktur der Daten durch den Einsatz der nichtparametrischen Dichteschätzung aufzudecken ist. In dieser Arbeit wird angenommen, dass die Daten einer gewissen Wahrscheinlichkeitsverteilung mit Dichtefunktion f unterliegen und zu den Domänen der Modi in f gehören.

Die Auswahl des Glättungsparameters hat einen großen Einfluss auf Kerndichteschätzung. In Kapitel 2 dieser Arbeit wurden vier Bandbreitenselektoren (LCV, LSCV, BCV und Direct-Plug-In) im uni- und multivariaten Fall anhand simulierter Daten verglichen, um das Verständnis der Rolle des Glättungsparameters bei Kerndichteschätzung zu vermitteln. Es hat sich ergeben, dass kein Bandbreitenselektor einen sowohl theoretisch als auch praktisch zur Kerndichteschätzung gut passenden Glättungsparameter liefern kann. Zum Zweck der Datenexploration ist zu empfehlen, dass man die unbekannte wahre Dichtefunktion durch Kerndichteschätzung mit einer Reihe von Glättungsparametern schätzt und die Eigenschaften der wahren Dichtefunktion anhand dieser Kerndichteschätzer untersucht, weil „different useful information can be available at different levels of smoothing.“

In manchen Fällen kann die wahre Dichte durch einen Kerndichteschätzer mit festem Glättungsparameter nicht gut widerspiegelt werden, insbesondere wenn die Daten einen langen Schwanz oder eine Multimodal-Struktur haben. In der multivariaten Datenanalyse ist diese Situation noch schlimmer, weil sie stark unter „Curse of Dimensionality“ leidet. Es wurde in der Literatur (S. 202 von Scott (1992), S. 90 von Wand & Jones (1995)) erwähnt, dass ein Kerndichteschätzer mit festem Glättungsparameter für die Dichteschätzung in den Fällen nicht geeignet ist, wenn die Dimension der Daten größer 5 ist. In Kapitel 3 dieser Arbeit wurden die folgenden zwei Anwendungsmöglichkeiten der Fest-Kerndichteschätzung in der multivariaten Datenanalyse diskutiert:

1. Man verwendet einen Fest-Kerndichteschätzer als Pilot-Dichteschätzer für das Kon-

struieren eines gemischten Modells, das besonders für die Daten mit einer Multimodal-Struktur gut geeignet ist;

2. Das auf einem Fest-Kerndichteschätzer basierte Clusteringverfahren liefert ein gutes Resultat bei der Untersuchung der unbekannten Modal- bzw. Cluster-Struktur in den hochdimensionalen Daten.

In Abschnitt 3.1 wurde ein Fest-Kerndichteschätzer als ein „Overparametrized“ Modell betrachtet und weiter als Pilot-Dichteschätzer für die Konstruktion eines gemischten Modells **Modell 4** verwendet. **Modell 4** wird dadurch konstruiert, dass man zuerst die Daten zu den Modi des Pilot-Dichteschätzers zuordnet und dann einen modifizierten SEM-Algorithmus darauf durchführt. **Modell 4** stellt einen Kompromiss dar, das in der Situation verwendet werden kann, wenn keine sichere Modellannahme vorhanden ist und ein Fest-Kerndichteschätzer scheitert.

In der explorativen Datenanalyse versucht man mit systematischer Anwendung von Methoden auf den Datenbestand, um die unbekannte Datenstruktur anhand der empirischen Daten aufzudecken. Im multivariaten Fall bezieht sich die Datenstruktur in erster Linie auf die Modi der Daten. Eine der wichtigsten Anwendungen der Dichteschätzung in der explorativen Datenanalyse liegt in der Identifizierung der Modalstruktur (Mode Hunting), was dann auf eine natürliche Weise zum Clustering der Daten führt. In Abschnitt 3.3 und 4.3 wurde anhand der simulierten und praktischen Datensätze gezeigt, dass die unbekannte Struktur der Daten durch die Zusammenarbeit des Prunings eines Dichteschätzer basierten Single oder Complete Linkage Dendrogramms und einer Visualisierungsmethode **Methode2** aufgedeckt werden kann.

Die Rolle der Datenvisualisierung in der explorativen Datenanalyse ist mit der weiteren Entwicklung der Computertechnik immer wichtiger geworden. In Kapitel 4 wurden verschiedene Visualisierungsmethoden in der multivariaten Datenanalyse anhand von Beispielen vorgestellt. Beim Anwenden der nichtparametrischen Dichteschätzung in der Datenexploration besteht die Notwendigkeit, die Information aus Kerndichteschätzung bzw. den darauf basierten statistischen Modellen zu visualisieren. Es wurde diskutiert, wie man die Information aus Kerndichteschätzung bzw. den darauf basierten statistischen Modellen mit in der Datenvisualisierung effektiv einbeziehen kann. Zwei Vorschläge für die Visualisierung der Information aus dem Dichteschätzer basierten hierarchischen Verfahren wurden in Abschnitt 4.3 gegeben, die sind, Visualisierungsbaum und **Methode2**. Es hat sich ergeben, dass diese zwei Visualisierungsmethoden beim Untersuchen der Clusterstruktur der Daten hilfreich sind.

In dieser Arbeit wurde die Anwendung der multivariaten Dichteschätzung in der explorativen Datenanalyse anhand simulierter und praktischer Datensätze diskutiert. Zusammenfassend sind dabei folgende Punkte zu beachten:

1. Bei Kerndichteschätzung sollen verschiedene Glättungsparameter genommen werden, weil „different useful information can be available at different levels of smoo-

thing;“

2. Unter der Annahme, dass die Daten einer Wahrscheinlichkeitsverteilung mit Dichtefunktion f unterliegen und zu den Domänen der Modi in f gehören, kann ein gemischtes Modell auf Basis eines Fest-Kerndichteschätzers (Pilot-Dichteschätzer) konstruiert werden, indem man die Komponenten des Pilot-Dichteschätzers durch Nutzung einer Variante des EM-Algorithmuses fusioniert. Dieses gemischte Modell hängt aber stark von dem Pilot-Dichteschätzer ab;

3. Die Datenstruktur im multivariaten Fall ist kompliziert und die Form bzw. das Wahrscheinlichkeitsmaß eines High Density Clusters ist schwierig zu beschreiben. Es wurde in Abschnitt 3.3 und 4.3 gezeigt, dass man die Clusterstruktur der multivariaten Daten durch Untersuchung der Baumstruktur des entsprechenden Dichteschätzer basierten Single bzw. Complete Linkage Dendrogramms aufdecken kann. Diese Baumstruktur hängt aber auch von dem darunter liegenden Dichteschätzer ab;

4. Aufgrund von (1), (2) und (3) ist zu empfehlen, die Information aus Kerndichteschätzung mit verschiedenen Glättungsparametern und darauf basierten statistischen Modellen zu visualisieren, damit man die Datenstruktur vollständig untersuchen kann.

In Zukunft sind die folgenden Punkte in Bezug auf diese Arbeit noch zu erforschen:

- Die Form bzw. das Wahrscheinlichkeitsmaß eines High Density Clusters ist noch genauer zu beschreiben (vgl. auch (3.18), (3.19) und (3.27));

- Neue Kriterien sind zu entwickeln, anhand denen die Glättungsparameter bei Kerndichteschätzung für das darauf basierte gemischte Modell bzw. Clustering auf eine automatische Weise bestimmt werden können;

- Neue statistische Verfahren sind zu entwickeln, um die Komponenten eines multivariaten Kerndichteschätzers zu fusionieren (vgl. Abs. 3.1 und Scott & Szewczyk (2001));

- In Abschnitt 3.3 und 4.3 dieser Arbeit wurde gezeigt, dass die Clusterstruktur in multivariaten Daten durch Untersuchung des Dichteschätzer basierten Baums aufgedeckt werden kann. Der Zusammenhang von Clustering und Graph ist noch weiter zu erforschen.

Literaturverzeichnis

[1] Agrawal, R., Gehrke, J., Gunopulos, D. and Raghaven, P. (2005). Automatic Subspace Clustering of High Dimensional Data. *Data Mining and Knowledge Discovery* **11**(1): 5-33.

[2] Aitchison, J. und C. G. G. Aitken. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63**(3): 413-420.

[3] Ankerst, M., Breunig, M. M., Kriegel, H-P. and Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure. *SIGMOD Rec.* **28**(2): 49-60.

[4] Berlinet, A. and Devroye, L. (1994). A comparison of kernel density estimates. *Publ. Inst. Statist. Univ. Paris*. **38**: 3-59.

[5] Bhattacharyya, D. K. and Borah, B. (2008). DDSC: A Density Differentiated Spatial Clustering Technique. *Journal of Computers* **3**(2): 72-79.

[6] Biau, G., L. Devroye and G. Lugosi. (2007). On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory* **54**: 781-790.

[7] Bouezmarni, T. and Rombouts, J. V. K. (2007). Nonparametric density estimation for multivariate bounded data. *CORE Discussion Paper 2007/65*.

[8] Brown, D. L., Cai, T. T. and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science* **16**(2): 101-133.

[9] Burman, P. and Polonik, W. (2009). Multivariate Mode Hunting: Data Analytic Tools with Measures of Significance. *Journal of Multivariate Analysis* **100**(6): 1198-1218.

[10] Celeux, G. and Govaert, G. (1992). A Classification EM Algorithm for Clustering and Two Stochastic Versions. *Computational Statistics and Data Analysis* **14**: 315-332.

[11] Celeux, G. and Govaert, G. (1995). Gaussian Parsimonious Clustering Models. *Pattern Recognition* **28**: 781-793.

[12] Chambers, J. M. (2008). *Software for Data Analysis: Programming with R*. New York: Springer.

[13] Chaudhuri, P. and J. S. Marron. (1999). SiZer for exploration of structure in curves. *Journal of the American Statistical Association* **94**: 807-823.

[14] Cheng, M. Y., Fan, J. and Marron, J. S. (1997). On automatic boundary corrections. *The Annals of Statistics* **25**(4): 1691-1708.

[15] Chiu, S-T. (2000). Boundary adjusted density estimation and bandwidth selection. *Statistica Sinica* **10**(2000): 1345-1367.

[16] Cleveland, W. S. (1993). *Visualizing Data*, Summit, NJ: Hobart Press

[17] Cleveland, W. S. (1994). *The Elements of Graphing Data*, Murray Hill, NJ: AT&T Bell Laboratories.

[18] Cook, D. and A. Buja. (1997). Manual controls for high-dimensional data projections. *Journal of Computational and Graphical Statistics* **6**(4): 464-480.

[19] Cook, D., A. Buja and J. Cabrera. (1993). Projection pursuit indices based on orthonormal function expansions. *Journal of Computational and Graphical Statistics* **2**(3): 225-250.

[20] Cook, D., A. Buja, J. Cabrera and C. Hurley. (1995). Grand Tour and Projection Pursuit. *Journal of Computational and Graphical Statistics* **4**(3): 155-172.

[21] Cook, D. and Swayne, D. F. (2007). *Interactive and Dynamic Graphics for Data Analysis: With Examples Using R and GGobi*. Berlin: Springer.

[22] Cox, D. R. and N. J. H. Small. (1978). Testing Multivariate Normality. *Biometrika* **65**(2): 263-272.

[23] Devroye, L. (1987). *A Course In Density Estimation*. Boston: Birkhäuser Verlag.

[24] Devroye, L. (1979). Recursive Estimation of the Mode of a Multivariate Density. *The Canadian Journal of Statistics* **7**(2): 159-167.

[25] Devroye, L. (1997). Universal smoothing factor selection in density estimation: theory and practice (with discussion). *Test* **6**: 223-320.

[26] Devroye, L. and Krzyzak, A. (1999). On the Hilbert kernel density estimate. *Statistics and Probability Letters* **44**: 299-308.

[27] Devroye, L. and Krzyzak, A. (2002). New Multivariate Product Density Estimators. *Journal of Multivariate Analysis* **82**(1): 88-110.

[28] Devroye. L. and Lugosi, G. (1996). A universally acceptable smoothing factor for kernel density estimation. *The Annals of Statistics* **24**: 2499-2512.

[29] Devroye, L. and Györfi, L. (1985). *Nonparametric density estimation: The L1 view*. New York: Wiley.

[30] Devroye. L. and Lugosi, G. (2000). *Combinatorial Methods in Density Estimation*. New York: Springer.

[31] Dong, J. P. and Simonoff, J. S. (1994). The Construction and Properties of Boundary Kernels for Smoothing Sparse Multinomials. *Journal of Computational and Graphical Statistics* **3**(1): 57-66.

[32] Duong, T. and Hazelton, M. L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics* **15**: 17-30.

[33] Duong, T. (2004). Bandwidth selectors for multivariate kernel density estimation. *Doctor's Thesis*. School of Mathematics and Statistics, University of Western Australia.

[34] Duong, T. and Hazelton, M. L. (2003). Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *Journal of Multivariate Analysis* **93**(2): 417-433.

[35] Duong, T. and Hazelton, M. L. (2005). Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation. *Scandinavian journal of statistics* **32**(3): 485-506.

[36] Duong, T., Cowling, A., Koch, I. and Wand, M.P. (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics and Data Analysis* **52**(9): 4225-4242.

[37] Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association* **90**(430): 577-588.

[38] Ester, M., Kriegel, H-P., Sander, J. and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In Proceedings of Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, 1996*: 226-231.

[39] Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**(2): 209-230.

[40] Fraley, C. and A. E. Raftery. (2002). Model-Based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* **97**: 611-631.

[41] Fraley, C and Raftery, A. E. (2006). Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. *Technical Report no. 486*. Department of Statistics, University of Washington. URL: http://stat.washington.edu/research/reports/2005/tr486.pdf

[42] Godtliebsen, F., Marron, J.S. and Chaudhuri, P. (2002). Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics* **11**: 1-22.

[43] Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association* **82**(397): 249-266.

[44] Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers* **C23**(9): 881-890.

[45] Friedman, J. H. and Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing* **9**(2): 123-143.

[46] Friedman, J. H. and Stuetzle, W. (2002). John W. Tukey's Work on Interactive Graphics. *The Annals of Statistics* **30**(6): 1629-1639.

[47] Friedman, J. H. and Meulman, J. (2004). Clustering Objects on Subsets of Attributes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**(4): 815-849.

[48] Geman, S. and Hwang, C. R. (1982). Nonparametric Maximum Likelihood Estimation by the Method of Sieves. *The Annals of Statistics* **10**(2): 401-414.

[49] Hall, P. (1987). On Kullback-Leibler Loss and Density Estimation. *The Annals of Statistics* **15**(4): 1491-1519.

[50] Hall, P. (1989). On polynomial-based projection indices for exploratory projection pursuit. *The Annals of Statistics* **17**(2): 589-605.

[51] Hall, P., Sheather, S. J., Jones, M. C. and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78**: 263-269.

[52] Hartigan, J. A. (1981). Consistency of Single Linkage for High-Density Clusters. *Journal of the American Statistical Association* **76**(374): 388-394.

[53] Hartigan, J. A. and Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics* **13**(1): 70-84.

[54] Hinneburg, A. (2003). Density-Based Clustering in large Databases using Projections and Visualizations. *Doctor's thesis*. Mathematisch-Naturwissenschaftlich-Technischen Fakultät der Martin-Luther-Universität Halle-Wittenberg.

[55] Hinneburg, A. and Gabriel, H-H. (2007). DENCLUE 2.0: Fast Clustering Based on Kernel Density Estimation. *Advances in Intelligent Data Analysis VII (2007)* **4723**: 70-80.

[56] Hjort, N. L. and Jones, M. C. (1996). Locally Parametric Nonparametric Density Estimation. *The Annals of Statistics* **24**(4): 1619-1647.

[57] Huber, P. J. (1985). Projection pursuit (with discussion). *The Annals of Statistics* **13**: 435-475.

[58] Hwang, J. N., Lay, S. R. and Lippman, A (1994). Nonparametric multivariate density estimation: a comparative study. *IEEE Transactions on Signal Processing* **42**(10): 2795-2810.

[59] Inselberg, A. (1985). The plane with Parallel Coordinates. *The Visual Computer* **1**: 69-91.

[60] Jones, M. C. and Sibson, R. (1987). What is projection pursuit? (with discussion). *Journal of the Royal Statistical Society: Series A* **150**: 1-36.

[61] Jones, M. C. and Foster, P. J. (1996). A simple nonnegative boundary correction method for kernel density estimation. *Statistica Sinica* **6**(1996): 1005-1013.

[62] Jones, M. C. and Henderson, D. A. (2009). Maximum likelihood kernel density estimation: on the potential of convolution sieves. *Computational Statistics and Data Analysis* **53**(10): 3726-3733.

[63] Klemelä, J. (2004). Visualization of multivariate density estimates with level set trees. *Journal of Computational and Graphical Statistics* **13**(3): 599-620.

[64] Klemelä, J. (2006). Visualization of multivariate density estimates with shape trees. *Journal of Computational and Graphical Satistics* **15**(3): 372-397.

[65] Klemelä, J. (2007). Visualization of multivariate data with tail trees. *Information Visualization* **6**: 109-122.

[66] Koo, J. Y. and Kooperberg, C. (2000). Logspline density estimation for binned data. *Statistics and Probability Letters* **46**: 133-147.

[67] Kooperberg, C. and Stone C. J. (1992). Logspline Density Estimation for Censored Data. *Journal of Computational and Graphical Statistics* **1**(4): 301-328.

[68] Ledl, T. (2004). Kernel Density Estimation: Theory and Application in Discriminant Analysis. *Austrian Journal of Statistics* **33**(3): 267-279.

[69] Lehmann, E. L. (2006). On likelihood ratio tests. *IMS Lecture Notes-Monograph Series 2nd Lehmann Symposium - Optimality* **49**(2006): 1-8.

[70] Li, J. (2005). Clustering Based on a Multi-layer Mixture Model. *Journal of Computational and Graphical Statistics* **14**(3): 547-568.

[71] Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics*, Princeton: Princeton University Press.

[72] Li, R. (2005). Local Likelihood SiZer Map. *The Indian Journal of Statistics* **67**(3): 476-498.

[73] Ligges, U. (2007). **Programmieren mit R**. Heidelberg: Springer.

[74] Loader, C. R. (1996). Local likelihood density estimation. *The Annals of Statistics* **24**(4): 1602-1618.

[75] Loader, C. R. (1999). Bandwidth Selection: Classical or Plug-In? *The Annals of Statistics* **27**(2): 415-438.

[76] Loh, W. L. and Zhang, C. H. (1996). Global properties of kernel estimators for mixing densities in discrete exponential family models. *Statistica Sinica* **6**: 561-578.

[77] Marchette, D. J., Priebe, C. E., Rogers, G. W and Solka, J. L. (1998). Filtered Kernel Density Estimation. *Computational Statistics* **11**: 95-112.

[78] Marron, J. S., Jones, M. C. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91**: 401-407.

[79] McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering.* New York: Marcel Dekker.

[80] McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: Wiley.

[81] Meila, M. (1999). Learning Mixtures of Trees. *Doctor's thesis*. Massachusetts Institute of Technology.

[82] Miasnikov, A. D, Rome, J. E. and Haralick R. M. (2004). A Hierarchical Projection Pursuit Clustering Algorithm. *17th International Conference on Pattern Recognition (ICPR'04)* **Volume 1**: 268-271.

[83] Michailidis, G. (2008). Data Visualization Through Their Graph Representations. *Handbook of Data Visualization*. Heidelberg: Springer.

[84] Minnotte, M. C. (1997). Nonparametric Testing of the Existence of Modes. *The Annals of Statistics* **25**(4): 1646-1660.

[85] Minnotte, M. C., Sain, S. R. and Scott, D. W. (2008). Multivariate Visualization by Density Estimation. *Handbook of Data Visualization*. Heidelberg: Springer.

[86] Müller, H. G. and Stadtmüller, U. (1999). Multivariate boundary kernels and a continuous least squares principle. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(2): 439-458.

[87] Nabney, l. T. (2002). *Netlab: Algorithms for Pattern Recognition*. London: Springer.

[88] Nason, G. P. (1992). Design and choice of projection indices. *Doctor's thesis*. University of Bath.

[89] Nason, G. P. (1995). Three-Dimensional Projection Pursuit. *Applied Statistics* **44**(4): 411-430.

[90] Oh, M. S. and Raftery, A. E. (2003). Model-based Clustering with Dissimilarities: A Bayesian Approach. *Technical Report no. 441*. Department of Statistics, University of Washington. URL: http://stat.washington.edu/www/research/reports/2003/tr441.pdf

[91] Priebe, C. E. (1994). Adaptive Mixtures. *Journal of the American Statistical Association* **89**(427): 796-806.

[92] Qin, Y. S. and Smith, B. (2004). Likelihood Ratio Test for Homogeneity in Normal Mixtures in the Presence of a Structural Parameter. *Statistica Sinica* **14**(2004): 1165-1177.

[93] Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge, U. K.: Cambridge University Press.

[94] Roeder, K. and Wasserman L. (1995). Practical Bayesian Density Estimation using Mixtures of Normals. *Journal of the American Statistical Association* **92**: 894-902.

[95] Roueff, F. and Rydén, T. (2005). Non-parametric estimation of mixing densities for discrete distributions. *The Annals of Statistics* **33**(5): 2066-2106.

[96] Rusu, A. and Santiago, C. (2008). Grid Drawings of Binary Trees: An Experimental Study. *Journal of Graph Algorithms and Applications* **12**(2): 131-195.

[97] Sain, S. R., Baggerly, K. A. and Scott, D. W. (1994). Cross-Validation of Multivariate Densities. *Journal of the American Statistical Association* **89**(1994): 807-817.

[98] Sain, S. R. (1994). Adaptive Kernel Density Estimation. *Doctor's Thesis*. Rice University.

[99] Sain, S. R. (2002). Multivariate locally adaptive density estimation. *Computational Statistics and Data Analysis* **39**: 165-186.

[100] Sander, J., Ester, M., Kriegel, H-P. and Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery* **2**(2): 169-194.

[101] Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*, New York: Wiley.

[102] Scott, D. W. and Wand, M. P. (1991). Feasibility of multivariate density estimates. *Biometrika*(1991), **78**(1): 197-205.

[103] Scott, D. W. and Sain, S. R. (2004). Multi-dimensional Density Estimation. *In Handbook of Statistics* **23**: *Data Mining and Computational Statistics (2004)*: 229-263.

[104] Scott, D. W. and Szewczyk, W. F. (2001). From Kernels to Mixtures. *Technometrics* **43**(3): 323-335.

[105] Selan Rodrigues dos Santos (2004). A Framework for the Visualization of Multidimensional and Multivariate Data. *Doctor's Thesis*. School of Computing, The University of Leeds.

[106] Silverman, B. W. (1981). Using Kernel Density Estimates to Investigate Multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)* **43**(1): 97-99.

[107] Silverman, B. W. (1982). Kernel Density Estimation Using the Fast Fourier Transform. *Applied Statistics* **31**(1): 93-99.

[108] Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman and Hall.

[109] Simonoff, J. S. (1996). *Smoothing methods in statistics*, New York: Springer Series in Statistics.

[110] Simonoff, J. S. (1998). Three Sides of Smoothing: Categorical Data Smoothing, Nonparametric Regression, and Density Estimation. *International Statistical Review* **66**(2): 137-156.

[111] Stuetzle, W. (2003). Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification* **20**: 25-47.

[112] Stuetzle, W. and Nugent, R. (2007). A generalized single linkage method for estimating the cluster tree of a density. *Technical Report 514*, Department of Statistics, University of Washington.

[113] Swayne, D. F., Lang, D. T., Buja, A. and Cook, D. (2003). GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics and Data Analysis* **43**(4): 423-444.

[114] Tantrum, J., Murua, A. and Stuetzle, W. (2004). Hierarchical model-based clustering of large datasets through fractionation and refractionation. *Information System* **29**(4): 315-326.

[115] Terrell, G. R. and Scott, D. W. (1992). Variable Kernel Density Estimation. *The Annals of Statistics* **20**(3): 1236-1265.

[116] Theus, M. (2002). Interactive Data Visualization using Mondrian. *Journal of Statistical Software* **7**(11).

[117] Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2): 411-423.

[118] Unwin, A. R. (1999). Requirements for Interactive Graphics Software for Exploratory Data Analysis. *Computational Statistics* **14**: 7-22.

[119] Unwin, A. R., Volinsky, C. and Winkler, S. (2003). Parallel coordinates for exploratory modelling analysis. *Computational Statistics and Data Analysis* **43**(4): 553-564.

[120] Unwin, A. R., Theus, M. and Hofmann, H. (2006). *Graphics of Large Datasets: Visualizing a Million*, Berlin/Heidelberg: Springer Series in Statistics and Computing.

[121] Urbanek, S. (2003). Many Faces of a Tree. *In Proc. of the 35th Symposium on the Interface of Computing Science and Statistics* **35**. Interface Foundation.

[122] Urbanek, S. (2008). Visualizing Trees and Forests. *Handbook of Data Visualization*. Heidelberg: Springer.

[123] Von Luxburg, U. and Ben-David, S. (2005). Towards a statistical theory of clustering. *In PASCAL Workshop on Statistics and Optimization of Clustering*.

[124] Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*, London: Chapman and Hall.

[125] Ward, M. O. (2008). Multivariate Data Glyphs: Principles and Practice. *Handbook of Data Visualization*. Heidelberg: Springer.

[126] Wegman, E. J. and Solka, J. L. (2002). On some mathematics for visualizing high dimensional data. *The Indian Journal of Statistics*. **64**(2): 429-452.

[127] Wilkinson, L. (2005). *The Grammar of Graphics*. Chicago: Springer Series in Statistics and Computing.

[128] Wilkinson, L. (2008). Graph-theoretic Graphics. *Handbook of Data Visualization*. Heidelberg: Springer.

[129] Wojciechowski, W.C. and Scott, D.W. (2000). High-Dimensional Visualization Using Continuous Conditioning. *Computing Science and Statistics* **32**: 267-274.

[130] Wong, M. A. (1982). A Hybrid Clustering Method for Identifying High-Density Clusters. *Journal of the American Statistical Association* **77**(380): 841-847.

[131] Wong, M. A. and Lane, T. (1983). A kth Nearest Neighbour Clustering Procedure. *Journal of the Royal Statistical Society. Series B (Methodological)* **45**(3): 362-368.

[132] Wu, H-M., Tzeng, S-L. and Chen, C-H. (2008). Matrix Visualization. *Handbook of Data Visualization*. Heidelberg: Springer.

[133] Zhang, C. H. (1990). Fourier Methods for Estimating Mixing Densities and Distributions. *The Annals of Statistics* **18**(2): 806-831.

[134] Zhang, C. H. (1995). On estimating mixing densities in discrete exponential family models. *The Annals of Statistics* **23**: 929-945.

[135] Zhang, C. H. (2005). Estimation of SUMS of Random Variables: Examples and Information Bounds. *The Annals of Statistics* **33**(5): 2022-2041.

[136] Zizek, F. (1938). Die verschiedenen Begriffe von „Statistik“. *Review of the International Statistical Institute* **6**(4): 519-552.

Curriculum Vitae

Person

- Aged 35. Married with one daughter
- China Citizen.

Education

- Ph.D. Mathematics, Universität Augsburg, 2010.
- Diplom Math. oec. Mathematics, Universität Augsburg, 2006.
- B.S. Transportation Engineering & Management, Dalian Maritime University, 1995.